Visual Content Characterization Based on Encoding Rate-Distortion Analysis

by

Zhuoran Li

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, 2023

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:	Jiying Zhao Professor, School of Electrical Engineering and Computer Science, University of Ottawa
Supervisor(s):	Zhou Wang Professor, Dept. of Electrical and Computer Engineering, University of Waterloo
Internal Member:	Otman Basir Professor, Dept. of Electrical and Computer Engineering, University of Waterloo
Internal Member:	Afred Yu Professor, Dept. of Electrical and Computer Engineering, University of Waterloo
Internal-External Member:	Giang Tran Assistant Professor, Dept. of Applied Mathematics, University of Waterloo

Author's Declaration

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

Visual content characterization is a fundamentally important but under exploited step in dataset construction, which is essential in solving many image processing and computer vision problems. In the era of machine learning, this has become ever more important, because with the explosion of image and video content nowadays, scrutinizing all potential content is impossible and source content selection has become increasingly difficult. In particular, in the area of image/video coding and quality assessment, it is highly desirable to characterize/select source content and subsequently construct image/video datasets that demonstrate strong representativeness and diversity of the visual world, such that the visual coding and quality assessment methods developed from and validated using such datasets exhibit strong generalizability.

Encoding Rate-Distortion (RD) analysis is essential for many multimedia applications. Examples of applications that explicitly use RD analysis include image encoder RD optimization, video quality assessment (VQA), and Quality of Experience (QoE) optimization of streaming videos etc. However, encoding RD analysis has not been well investigated in the context of visual content characterization. This thesis focuses on applying encoding RD analysis as a visual source content characterization method with image/video coding and quality assessment applications in mind. We first conduct a video quality subjective evaluation experiment for state-of-the-art video encoder performance analysis and comparison, where our observations reveal severe problems that motivate the needs of better source content characterization and selection methods. Then the effectiveness of RD analysis in visual source content characterization is demonstrated through a proposed quality control mechanism for video coding by eigen analysis in the space of General Quality Parameter (GQP) functions. Finally, by combining encoding RD analysis with submodular set function optimization, we propose a novel method for automating the process of representative source content selection, which helps boost the RD performance of visual encoders trained with the selected visual contents.

Acknowledgements

First and foremost, I would like to express my sincere thanks to my supervisor Prof. Zhou Wang. I feel tremendously fortunate to be accepted as his graduate student in the Fall of 2017. During my 5-year PhD journey, I received profound and valuable guidance from Prof. Wang through the countless research discussions with him. He taught me what is valuable research and the skills on the art of writings and presentations, which would have a long-lasting beneficial impact on my future professional career. His professional attitude and continuous enthusiasm towards the research has always been the virtues that enlightened and inspired me. Without his supportive guidance and our team's collaborative environment cultivated by him, I would not have the accomplishments in this PhD thesis. Moreover, I am honored to have Dr. Jiying Zhao, Dr. Otman Basir, Dr. Afred Yu and Dr. Giang Tran on my thesis committee.

Secondly, I would like to express sincere thanks to the talented members of my research team, the Image and Vision Computing (IVC) team. These include Zhengfang Duanmu, Wentao Liu, Shahrukh Athar, Rasoul Mohammadi Nasir, Kai Zeng, Abdul Rehman, Jiheng Wang, Hojatollah Yeganeh, Zhongling Wang, Jinghan Zhou, Xiaoyu Xu, Sheyang Tang, Mahdi Naseri, Armin Shafiee Sarvestani, Mahzar Eisapour, and Paul Yang. Particularly, Zhengfang Duanmu and Wentao Liu helped me a lot as senior members of the team on both academic research and life outside the lab. They are my supportive colleagues and close friends.

Finally, I wish to extend my gratitude to my family, including both my parents and my wife. My parents, Yueying Zhang and Qiang Li, have always been the most supportive people for my academic career starting from my undergraduate career as an international student in Canada. It is their unconditional love that makes me achieving my goal of studying as a PhD student and needless to consider burdens as an international student. Moreover, words cannot express how much I appreciate the support in life I get during my PhD career from my wife, Tong Qi, who is the one that always stand firmly behind me when I am facing challenges and took care of the whole family during my arm injury. It is their countless support that helped me to reach the final line as a PhD student.

Dedication

This is dedicated to my new born daughter, Evelyn Li.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

4K 3840 × 2160 or 4096 × 2160 pixel resolution 42–44, 46, 57, 111

ACR Absolute Category Scale 48

AV1 AOMedia Video 1 xiv, 25, 42–44, 46, 49–52, 57, 98, 115

AVC Advanced Video Coding xiv, 42–44, 46, 49, 51, 52, 57

AVIF AV1 Image File Format 98

AVS2 Audio Video Coding Standard 2 xiv, 42–44, 46, 49, 51, 52, 55, 57

CF Colourfulness 17, 56, 86, 88, 106–108

CNN convolutional neural network 113–115

CQP Constant Quantization Parameter 29

CRF constant rate factor 29, 30, 46, 53

CTU Coding Tree Unit 19, 27–29, 114

CU Coding Unit 49

DCT discrete cosine transform 21

DNN deep neural network 6, 9, 87, 112

DST discrete sine transform 21

E2E End-to-End xiv, 6–9, 31, 32, 39, 40, 60–64, 66, 81–83, 85, 87, 98, 99, 112

FHD Full High Definition 43, 44

FPS Frames per Second xi, 47

GDN Generalized Divisive Normalization 33

GOP Group of Pictures 46

GQP General Quality Parameter iv, 4, 7

GRD generalized rate distortion 18

HDR High Dynamic Range 88

HEIF High Efficiency Image File Format 98

HEVC High Efficiency Video Coding xiii, xiv, 7, 9, 18, 21–23, 25, 27–29, 31, 40, 42–44, 46, 49, 51, 52, 55, 57, 98, 115

HVS Human Visual System 2, 5, 10, 14, 29–31, 43, 57, 62, 89, 112

IC image complexity 16

IQA image quality assessment 2

ITU-R ITU Radiocommunication Sector 48

ITU-T ITU Telecommunication Standardization Sector 46

JPEG Joint Photographic Experts Group 98

MOS mean opinion score 17, 18, 48, 49, 53, 54

MSE Mean Square Error 31

NLP Natural Language Processing 6

PCA Principal Component Analysis 103, 106

PLCC Pearson linear correlation coefficient 48

PSNR Peak Signal to Noise Ratio 2, 31

QoE Quality of Experience iv, 43, 91

QP Quantization Parameter 22, 25, 28–30, 57, 60, 61, 63, 64

RD Rate-Distortion iv, xiii, xiv, 2–9, 17–19, 22, 25–34, 39, 40, 42, 49–55, 57–60, 63, 86, 87, 91, 93, 96, 98, 99, 101, 103, 105–108, 111–116, 118

RDO Rate-Distortion Optimization 27, 28, 32

RMSE root mean square error 16

SI Spatial Information xi, 2, 4, 5, 14, 15, 17, 18, 40, 44, 47, 53, 54, 56, 57, 86, 88, 93, 106–108, 112

SRCC Spearman rank-order correlation coefficient 48

STD standard deviation 48

TI Temporal Information xi, 2, 4, 5, 15, 17, 18, 40, 44, 47, 53, 54, 57

UHD Ultra High Definition 42–44

VQA video quality assessment iv, 2, 43, 57

Chapter 1

Introduction

1.1 Motivation

Visual content characterization, which can be defined as a description of the distinctive nature or features for visual content, is a concept frequently used in the database construction of many visual-related research areas. For instance in computer vision area, during the database construction, the visual content such as image or video is classified by human labellers through characterizing visual contents based on different semantic categories. However, unlike its direct application on visual content classification database, visual content characterization has not been deeply investigated in the source content selection process for image/video coding and visual quality assessment database. Visual quality database usually consists of source pristine contents, distorted contents generated from the source contents, and human subjective quality ratings. It is a common practice for researchers to make empirical use of some low-level characteristics such as brightness, contrast, and sharpness etc. to select source contents. Even though visual quality assessment is considered to be a low-level vision task, there is no proof that those low-level features can describe the source content accurately from visual coding and quality perspective.

The concept of visual complexity has been proposed to capture source visual contents' level of compression difficulty. [3] With the assumption that several visual traits, such as

edge information in image and temporal frame difference in video, may reflect the compression difficulty of source contents, the visual features of Spatial Information (SI) and Temporal Information (TI) have been proposed and included in the ITU recommendation on guiding the source content selection for visual quality database. [4] However, with the many visual characteristics at hand, it is hard, if not impossible, to verify how each of them affects source contents from quality perspective. Therefore, in image and video quality studies, researchers empirically combined the aforementioned low level traits, namely brightness, contrast, and sharpness etc., with the SI and TI on selecting source contents. In an ideal scenario, researchers should select source contents based on the aforementioned visual characteristics. However, in reality for most quality database studies, the visual characteristics are only used for demonstrating the diversity of selected source contents through drawing a convex hull on the scatter plot. The source contents are actually hand-picked by researchers based only on their subjective judgement, where the subjective bias across different researchers would greatly affect the source content selection process. This is largely due to the lack of reliable systematic methods to automate the selection process that guarantees the selected source contents is representative for specific applications. More details will be discussed further in the background chapter.

The research on Rate-Distortion (RD) analysis in visual quality area originates from modelling and predicting the RD behaviour for different source contents when they are compressed into different lossy compression levels [5, 6]. With an accurate RD modeling for source contents, image and video encoders are expected to select appropriate sets of coding parameters so that the least possible distortion, usually measured in Peak Signal to Noise Ratio (PSNR), can be achieved given the limited data rate budget, which is considered a better RD performance in the literature. In recent decades, with the success of Human Visual System (HVS) driven image quality assessment (IQA) and video quality assessment (VQA) models such as SSIM [7] and SSIMplus [8], it is possible to automate visual quality assessment process with high correlation against human perception of compression quality. The application of HVS driven quality models on RD analysis not only makes compression more meaningful in terms of human quality perception, but also opens the door for analyzing the visual source content characteristics from HVS perspective because the RD statistics can be directly linked to perceived quality by humans. Moreover,

unlike other visual characteristics that only capture source content characteristics from a single perspective using one number, the RD statistic is a high dimensional measure since one need to compress source contents into different lossy compression levels to obtain the RD curve, which can be further expanded into more than two dimensions when considering other factors such as resolution [9]. When compared to other visual characteristics, RD enjoys a much deeper relationship with quality by serving as a visual characteristics summarization measure by leveraging visual encoder as a source content analyzer.

With the compression quality related characterization measure at hand, the last piece of puzzle towards representative source content selection is the automatic objective content selection framework. In visual quality research area, source content selection usually starts with a large collection of pristine contents, which is essentially a subset selection process with each pristine content treated as an element, the collection of pristine contents treated as the ground set, and the selected source contents treated as a subset. Submodular set function, naturally modeling notions of information, diversity, and coverage in many applications [10], can be used to measure selected subset's representativeness. As proved by its application in many machine learning tasks [11], the optimization of submodular set functions that come with "diminishing return" property, is both theoretically sound and practically useful for solving subset selection problem. Therefore, the submodular set function optimization solution nicely fits the task of selecting source contents for compression quality database. Therefore, we solve the source content selection problem in RD domain by submodular set function optimization, which leads to a systematic objective source content selection framework that guarantees the representativeness.

1.2 Objectives

The primary objective of the thesis is to address the visual content characterization problem for image/video compression and quality related tasks by means of RD behavior analysis of visual content, leveraging visual encoders as an analytical tool. The necessity of effective visual content characterization, which is briefly discussed in the motivation section, is demonstrated through a thorough review of

the visual content characterization problem, with a specific focus on its applications for image/video compression and quality-related tasks. In the literature review at Chapter 2 and the subsequent quality subjective experiment at Chapter 3, widely used heuristic visual traits utilized to describe visual characteristics are proven to be unreliable, while the effectiveness of utilizing encoding RD behavior analysis to address the visual content characterization problem is validated.

Based on the fact that visual encoders can serve as content characteristics analyzers through encoding RD analysis, we aim to address two real-world engineering needs as sub-objectives of the thesis.

The first sub-objective is the precise perceptual quality control of image/video encoders. As discussed in the motivation section, over-simplified visual traits such as SI and TI are unreliable but play key roles in the visual encoder control process for characterizing source contents. This can lead to inferior quality control performance for modern visual encoders. By adopting the primary objective's "encoder as analyzer" philosophy, we propose the eigen-General Quality Parameter (GQP) method. This method is inspired by encoding RD modeling and effectively solves the long-standing problem of precise visual encoder perceptual quality control.

The second sub-objective is the source content selection in image quality database construction. As briefly mentioned in the previous section, the motivation originates from the need for a systematic objective visual quality database source content selection process. Currently, this process relies on the empirical use of low-level visual characteristics or even researchers' personal preferences based on their "expert experiences". To address this issue, we propose a systematic source content selection framework for compression and quality-related applications, using encoding RD statistics and submodular representativeness optimization. This framework is validated through a learning-based image compression framework and effectively replaces the unjustified low-level visual characteristics and easily biased "expert experiences" in content selection problems.

1.3 Contributions

Visual content lives in an extremely high dimensional space if each pixel value is treated as a single dimension. Moreover, the growing accessibility of image and video capturing devices such as smartphones results in an explosion of the amount of visual content in recent years. The high dimensionality and high volume of visual content make it impossible for all content to be scrutinized by humans in their life times, and thus it is imperative to smartly summarize and select from collected visual content, upon which visual quality and visual information compression research can be conducted. The visual characteristics ideally should be low dimensional and convenient for processing. Moreover, the visual characteristics, if properly designed, make it possible for researchers to objectively select source contents for the quality databases. **The first contribution in this thesis is to point out the importance of establishing visual content characterization as a critical problem in visual compression and quality research (as opposed to purely empirical content selection), and for the first time to provide a comprehensive review and analysis of the related literature.** Although previously several indicators have been used by researchers for selecting source contents, a comprehensive literature review on those indicators and their underlying methodology is lacking. The review of the visual content characterization problem provides a common ground for future compression and quality database related research.

As discussed in the motivation section, the characteristics such as brightness, contrast, SI, and TI etc have been widely used for selecting source visual contents while their effectiveness on capturing the source contents from quality perspective is unverified. The aforementioned characteristics are only collected from source pristine content, making it untenable to claim they are suitable for compression applications. **Therefore, the second contribution of the thesis is that we propose to use encoding RD statistics as a better visual content characterization method for compression quality related tasks.** Since the RD statistics can directly include HVS inspired quality models as measures to explain the distortion of lossy encoded contents, using RD as a characteristic for source visual contents is well motivated and promising. Moreover, we provide several real case examples to verify the usefulness of RD analysis in the background chapter and using

a specific application for RD analysis on encoding quality control. By applying the proposed RD analysis method, which is an eigen analysis approach, on video encoder control parameter against the encoded video quality, the proposed framework clearly outperforms the current quality control method and makes it possible to predict the relationship between quality and encoding parameters in a time efficient manner, precluding the necessity of designing ad-hoc methods aiming for saving encoding time. Moreover, the RD analysis inspired quality control method not only works on video encoders, but also deep neural network (DNN) driven End-to-End (E2E) image encoder.

Source visual content selection has always been a difficult task for visual quality dataset construction. Since the subjective experiment capacity is always limited, researchers can only test tens of source contents together with their distorted versions at most for human subjective quality evaluation in a lab setting. On the other hand, the pristine visual contents are so abundant, especially in recent years, thanks to the online content viewing platform such as YouTube and Netflix. Therefore, the challenge becomes obvious for quality research: how to select representative source contents given the uncountable large amount of pristine source visual contents? Nowadays, researchers usually start the database construction process by selecting a subset of data from a large collection of pristine visual contents. The selection procedure is usually subjective and biased since there is no objective selection procedure or testing criterion/tool available in the literature and the problem has often been neglected by researchers. Researchers have to select contents randomly or based on their own "expert experience". **In this thesis, the third contribution is that we provide a theoretically sound and practically useful framework for source content characterization and selection.** The third contribution is built based on the previously proposed quality related visual content characteristics, encoding RD statistics. Subset selection problem is frequently visited in machine learning area, especially for Natural Language Processing (NLP) tasks such as text summarization. Among the many solutions to subset selection problem, submodular set function optimization is an effective approach that has been proven both in theory and practice. Our proposed framework utilizes the submodular optimization approach.

1.4 Thesis Outline

The thesis is organized as the following:

In Chapter 2, visual characterization concept is introduced, followed by a review of frequently used visual characteristics in quality research area including their implementations and applications. Moreover, the background of encoding RD analysis is discussed in detail. The validity of using encoding RD statistics as a characterization measure is addressed. Lastly the modern video encoder High Efficiency Video Coding (HEVC) and recently successful neural network based E2E image compression model are described, which will be utilized later in the thesis.

In Chapter 3, we present our work on quality subjective experiment with the application on encoder comparison, which reveals the problems of visual content characterization and source content selection. Firstly in Section 3.1, the purpose of a comprehensive subjective and objective assessment of encoders is elaborated. Then the database construction and experiment setup are described thoroughly in Section 3.2. Lastly in Section 3.3 and Section 3.4, based on the subjective experiment results and the problems encountered during the database construction, the objectives of the thesis are analyzed.

In Chapter 4, the effectiveness of encoding RD analysis is demonstrated through its application in the precise control of visual coding. Section 4.1 gives an introduction of the related works and the connection between RD analysis and encoder quality control is drawn. In Section 4.2 and Section 4.3, the mathematical foundation is introduced and therefore the modelling of GQP is proposed. In Section 4.4, the eigen analysis guided algorithm for recovering GQP space from sparsely sampled data points is described. The effectiveness of the proposed framework is validated by the experiments in Section 4.5 and Section 4.6. Lastly in Section 4.7, the conclusions are drawn and further directions of research are discussed.

In Chapter 5, a novel source content selection framework based on submodular optimization in RD domain is proposed. Firstly in Section 5.1, the background of compression quality dataset construction and the benefit of selecting contents in RD domain is introduced. Secondly in Section 5.2, the submodular optimization and its suitability for the

problem of content selection are discussed in detail. In Section 5.3, the source content selection procedure is introduced. In Section 5.4, experiment using E2E learning based image compression model is conducted to validate the usefulness of the proposed measure in promoting the representativeness of image quality dataset for compression applications.

In Chapter 6, the thesis is summarized, the importance of including encoding RD analysis as a visual content characterization measure for compression applications is reemphasized. Future directions are discussed.

Chapter 2

Background

This chapter begins with the introduction of the visual content characterization concept from low level vision perspective, which plays a key role in visual quality research, especially in visual quality dataset construction. In the area of visual coding and visual quality, visual content characterization is a broad concept not only includes some common low level vision traits such as brightness, contrast, and sharpness, etc, but also other compression related factors such as visual complexity. The low level vision traits and visual complexity will be reviewed in the context of image and video quality database construction. Moreover, the usage of encoding Rate-Distortion (RD) performance statistics as a better characterization method for compression applications will be investigated through a review of its existing applications and observations about its capability in reflecting source content compression related characteristics.

The encoding RD inspired precise quality control proposed later in Chapter 4 is built upon the video and image encoders. Therefore, the second half of this chapter will review the modern video encoding pipeline using High Efficiency Video Coding (HEVC) as an example, specifically focusing on the rate and quality controlling mechanism. Moreover, the deep neural network (DNN) based End-to-End (E2E) image compression method using convolutional network and optimized in E2E manner will be reviewed [12]. The idea of DNN driven E2E learning incorporated into the encoding framework plays a key role in its state-of-the-art RD performance but at the same time brings major new challenges to

encoder control, hindering it from real applications.

2.1 Visual Content Characterization

2.1.1 Low level vision traits

In vision science, a hierarchical framework is typically considered as a way to explain human visual processing framework. It consists of a series of discrete stages that successively produce increasingly abstract representations. Shown in Fig.2.1, the stages are often considered in terms of low-, mid- and high-level representations. Low-level vision is related to representations of elementary features such as local colour, luminance, spatial frequency, orientation or contrast while high level stage is related to categorical or semantic representations to enable classification or identification [13]. Since the Human Visual System (HVS) quality assessment process does not involve classification or identification and human participants typically give out a quality rating based on their general perception during the visual quality subjective experiment, the quality assessment is believed to be mainly a vision task involving low-level vision features. Therefore, the features frequently used in vision science area for low-level vision perception are borrowed for visual quality assessment purposes, such as luminance (equivalent to brightness in the thesis), color, and contrast etc.

Brightness

Recommended in ITU-R BT.601 [14], brightness is defined as a weighted sum of the RGB three channels of source signal.

$$Y = 0.299r + 0.587g + 0.114b \tag{2.1}$$

In the image quality database CID2013 [15], brightness was used for the first time as a measure guiding the source content selection. The author claims that the brightness distribution of their selected source contents represent typical photographs that consumers

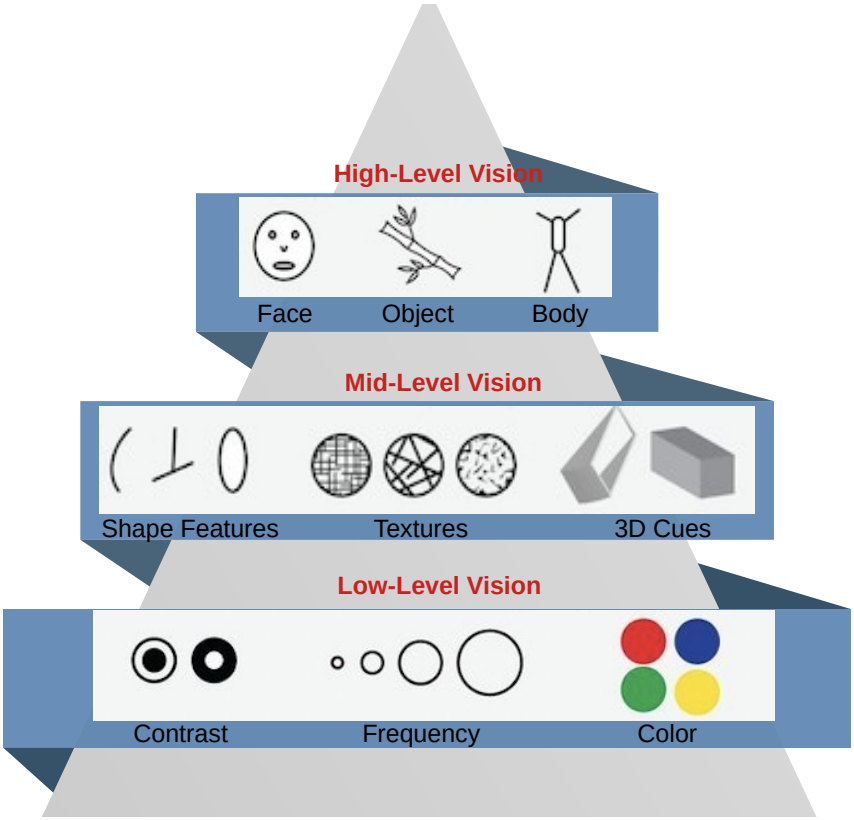


Figure 2.1: Hierarchical human vision system

might capture with their cameras. The source contents are selected based on the brightness levels aiming for a good coverage of typical photographs brightness range. In KonIQ-10K image quality dataset [16], around 10 thousand images are selected in a uniformly sampling strategy according to the brightness distribution of the 4 million image collection. It should be noted that the authors of KonIQ-10K database use the uniform sampling as a way to ensure the diversity of selected images. In case of video quality databases, the MCL-V streaming video quality database proposed in 2015 is the first work that takes brightness into consideration [17]. In this work, the brightness for source contents are divided into three levels and a summary of each content's brightness level is shown in a table demonstrating that the 12 source contents have a good coverage of brightness

levels. In YouTube UGC database [18], the brightness distribution is plotted against other databases and the convex hull is plotted for the 2 dimensional plot of brightness and contrast to demonstrate the diversity since its convex hull's coverage is the largest compared to other databases.

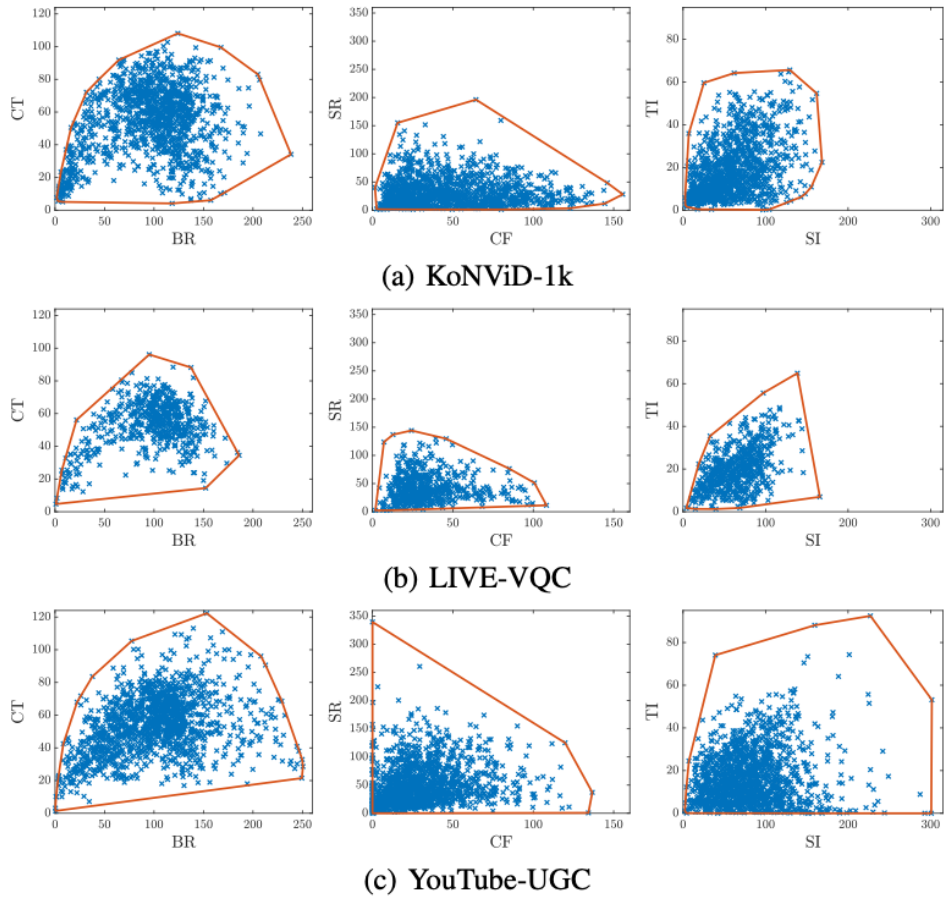


Figure 2.2: Scatter plots examples with six features: contrast (CT), brightness (BR), sharpness (SR), colorfulness (CF), temporal information (TI), spatial information (SI)

Contrast

Contrast, defined as the difference between the maximum brightness and the minimum brightness, is an indicator used by many quality datasets for selecting source contents as well. In authentically distorted image quality dataset KonIQ-10K [16], contrast is used together with other visual content characteristics such as brightness, colorfulness by uniformly drawing selected contents to ensure the diverse selection. For video quality database, the distribution of contrast for three database, KoNVid-1k [19], Live-VQC [20], and YouTube-UGC [18] are plotted and compared in the paper [21], where the authors use contrast together with other indicators to review the selected contents. In addition to distribution plot, the scatter plots of contrast versus brightness are drawn with the convex hull. The practice of drawing visual characteristics on a scatter plot with convex hull is common in visual quality database studies. An example is shown in Fig.2.2 [21].

Colorfulness

Colorfulness is another visual characteristics widely utilized in the quality database works. Proposed by Hasler et al [22], the measure is defined using the difference between R, G, and B channels:

$$rg = R - B \tag{2.2}$$

$$yb = \frac{1}{2}(R + G) - B \tag{2.3}$$

The standard deviation σ and the average μ of the color differences are further calculated and utilized in the final empirical colorfulness measure C.

$$\sigma_{rgyb} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} \tag{2.4}$$

$$\mu_{rgyb} = \sqrt{\mu_{rg}^2 + \mu_{yb}^2} \tag{2.5}$$

$$C = \sigma_{rgyb} + 0.3\mu_{rgyb} \tag{2.6}$$

For image datasets, colorfulness are usually combined with Spatial Information (SI) to demonstrate the diversity of source contents. Examples can be seen in the MDID dataset [23] and Waterloo Exploration-II dataset [2]. The authors in those image quality dataset papers draw a 2D convex hull of the scatter plots with the assumption that the diversity of selected contents is positively correlated with the area of the convex hull, though no justification of such an assumption has ever been provided. The colorfulness for each video frame is averaged for evaluating the video colorfulness in the BVI-CC video compression quality database [24]. The distribution of colorfulness for three database, KoNVid-1k [19], Live-VQC [20], and YouTube-UGC [18] are plotted and compared in the video quality database review paper [21].

2.1.2 Visual complexity measures

The need for visual complexity measures arises from the need of evaluating video compression related dataset using simple metrics.[3] Researchers want to find a metric that can capture the visual content's level of complexity viewed by HVS and therefore expect the level of complexity is related to the source content's level of compression difficulty. For this purpose, several measures have been proposed to evaluate visual complexity, which can be classified into filter based solutions and compression based solutions.

For filter based solution, it is assumed that the complexity can be explained in the level of edge contents.[25, 26, 27] Among many of the visual complexity measurements, the work by Yu et al.[3] proposed a Sobel filter based metric to evaluate image complexity, which is the recommended method SI in ITU recommendation based upon [4]. In the study, two complexity measures are proposed based on the Sobel filter. Let s_h and s_v denote gray-scale images filtered with horizontal and vertical Sobel kernels, respectively:

$$SI_r = \sqrt{s_h^2 + s_v^2} \tag{2.7}$$

Let P denote the number of pixels in the image. Two complexity measures are defined as

$$SI_{mean} = \frac{1}{P} \sum SI_r \tag{2.8}$$

$$SI_{stdev} = \sqrt{\frac{1}{P}\sum(SI_r - SI_{mean})^2} \tag{2.9}$$

Since the above methods can only evaluate image complexity or video frame-level complexity, the video complexity from temporal perspective in ITU recommendation [4] is defined as:

$$TI = max_{time}\{std_{space}[M_n(i,j)]\} \tag{2.10}$$

where $M_n(i,j)$ is the motion difference defined as the pixel value difference at location (i,j) between the two adjacent frames $F_n(i,j)$ and $F_{n-1}(i,j)$:

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j) \tag{2.11}$$

Regardless of the adoption of the filter-based method SI and frame difference based Temporal Information (TI) in ITU standard recommendation, they only capture limited aspects of visual contents, namely edge information and frame difference. What is more, their effectiveness is questionable on characterizing extremely complex visual content with just the two over-simplified single dimension features.

As will be discussed in the video encoding pipeline section, encoder control for lossy visual content is a hard problem. Because of the high dimensional data space, the encoding parameters have to be adapted based on different characteristics for the large amount of images or videos. Though visual contents can be characterized by contrast, color, texture, SI, or TI etc, the encoders often have to take additional visual characteristics and image statistics into account to find the best encoding parameters, which is a difficult problem, if not impossible, given our limited understanding of the visual content space. Therefore, in practice, SI and TI find limited usage on guiding encoding parameter selection for different source contents because they could not meet the need of accurately describe source contents from compression perspective for precise encoder control.

Compression based complexity measures are inspired by the concept of Komogrov complexity.[3] As stated in Komogrov complexity theory, an object (such as an image)'s complexity is represented by the length of the shortest computational program that can

describe the object. Due to the fact that Komogrov complexity is not computable, it can only be approximated by real-world compressors [28]. Several image complexity (IC) metrics based on the Komogrov complexity are listed as below:

$$IC_{LS} = \frac{1}{CR} \tag{2.12}$$

$$IC_{RMSE}(q) = \frac{RMSE(q)}{CR(q)} \tag{2.13}$$

$$IC_{LY}(q) = \frac{1}{CR(q)} \tag{2.14}$$

where CR denotes the compression ratio defined by $CR = \frac{s(I)}{s(C(I))}$, where $s(I)$ is the uncompressed image or video file size and $s(C(I))$ is the encoded bitstreams' size of the compressor C. When an image is losslessly compressed, the IC_{LS} represents the lossless image complexity. When the lossy compressed image file size is achieved at certain compression levels q, the image complexity $IC_{LY}(q)$ can be defined. In computing aesthetics [29, 30], lossy compression and distortion are used to define image complexity, where $RMSE$ is the root mean square error (RMSE) between the original image and the lossy compressed image, and q is a parameter that controls the level of lossy compression.

The compression based metrics are used as ground truth for some complexity metrics' performance assessment. In [3], correlation against the image compression based complexity measure is used as a way to validate the effectiveness of the proposed model. The evaluation process leads to a paradoxical situation: If Komogrov complexity motivated complexity measures are considered as alternatives to visual content complexity measures, then they are not supposed to be used as ground-truth in evaluating other filter-based metrics.

2.1.3 Visual complexity and human perception

Conceptually, how much a visual content can be compressed is related to the complexity of the content, which naturally brings about the problem of visual content complexity comparison. Many subjective experiments are conducted in reviewing the complexity level

of different content so as to evaluate the performance of their proposed complexity metrics [31, 27, 32]. In paper [32], a subjective experiment is conducted on evaluating video content complexity. A total of nine videos are watched by three participants. Viewers are asked to rate the level of video complexity in a 5 scales rating. The average scores are obtained after the subjective experiment. Since the number of participants is limited and no statistical analysis is done to verify the importance of obtained data, the experiment is only treated as preliminary. Recently in the work by Durmus [31], a more comprehensive image complexity subjective experiment is done. There are a total of 16 contents, ranging from natural scenes, paintings to simple lines. Forty-four viewers participate in the experiment and the final results show that mean opinion score (MOS) for image complexity varies significantly across participants, suggesting that subjective experiments on verifying the designed metrics are not conclusive. The author admitted that the observers' judgements should be "taken with a grain of salt".

As can be observed from the past subjective experiments on visual complexity, there is a large variation in human observers' opinions on visual complexity and it is often unrealistic to achieve a consensus on the complexity level for a specific video or image content. As a result, it may only be considered as both the filter based and compression based features visual content characteristics instead of visual complexity measure. Moreover, the subjective experiments suggest that human judgements are not reliable for evaluating the visual contents for compression applications. Consequently, during the source content selection for compression quality database, human bias towards different contents could significantly affect the performance of the compression applications utilizing the database.

In comparison with subjective visual complexity, visual content characterization may be a better concept to describe visual content for visual quality and compression tasks. Besides SI and TI, other visual content features such as Colourfulness (CF), contrast, and RD statistics may also be incorporated and a better description of visual content is highly desirable.

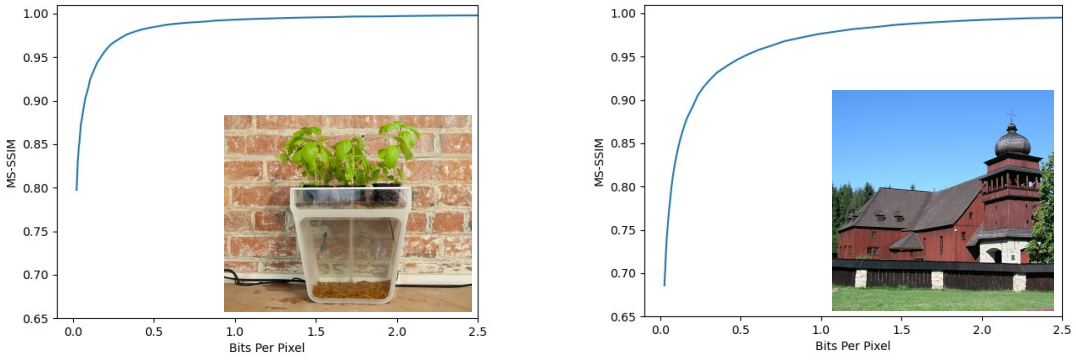
2.1.4 Encoding RD analysis as a characterization measure

Encoding RD analysis is an essential step in evaluating different data compressors' performance. In video and image encoding area, new generation encoder's improvement is validated through RD analysis, which is conducted by encoding source contents into a range of compression levels, so that RD curves can be drawn to compare quality improvement given the same data rate or rate saving at the same quality level [33, 34, 35, 36, 37]. Since different video contents may vary drastically in their RD curves, an aggregate evaluation of a large number of contents' RD performance offers a meaningful overall performance evaluation. Moreover, with researchers paying more attention to other factors affecting compression quality, such as spatial resolution and frame rate, the RD analysis may be extended to a generalized rate distortion (GRD) analysis.

The RD statistics contain more information than widely used characterization methods such as SI and TI. As seen in Fig.2.3, the two images share the same SI value 68 but the RD curve for the natural scene with building content is lower than that for the flowerpot content on the right, which indicates the flowerpot is actually easier to be encoded in terms of compressibility. For video content, the two contents in Fig.2.4 not only share the same SI value 54, but also have close TI values 23 and 24. However, their RD curves behave differently and cross each other. In this figure, the y axis represents the MOS value collected in the subjective experiment in Chapter 3.

Unlike the complexity measure such as SI that can be compared in terms of a single dimensional value only, the RD curves shown in Fig.2.3 and Fig.2.4 contain more information not just limited to the level of source content compressibility. As shown in Fig.2.5, the RD curve for the river content is higher than that of the flower content at low bitrate range while lower in the high bitrate range. The two videos are encoded using HEVC video encoder under 90 bitrate levels at 1080p resolution. The common crossing phenomenon in video RD analysis implies that single dimensional numbers are insufficient and richer information about visual content characteristics are captured by encoding RD analysis.

What is more, as pointed out in [38], visual content details are embedded in multiple resolutions that have varying impact on the perceptual quality evaluation of the contents. The RD statistics can be extended to incorporate RD information on different resolutions,



(a) (b)

Figure 2.3: Samples of RD curves for different image contents with the same SI

which are exemplified by 3-dimensional surfaces in Fig.2.6 [1].

Therefore, we propose to characterize visual content using encoding RD analysis for compression applications. In our proposal the visual encoders directly work as analyzers through the encoding RD analysis of lossy encoding. We intend to characterize visual content directly from compression perspective, which makes it better matched to compression and related quality evaluation tasks.

2.2 Video Encoding Pipeline

Fig 2.7[39] illustrates how typical predictive video encoder and decoder work. Firstly, the video frame is partitioned into image patches, which are called Coding Tree Unit (CTU). Then for each CTU, based on its characteristics, the encoder will either pass it through intra-frame estimation (a.k.a intra-prediction) or motion compensation (a.k.a inter-prediction). After all the predictions are done, the encoder subtracts the predicted signal from the original CTU signal, and the residual will go through the transform, scaling and quantization processes. The decoder reverses the aforementioned process to reconstruct

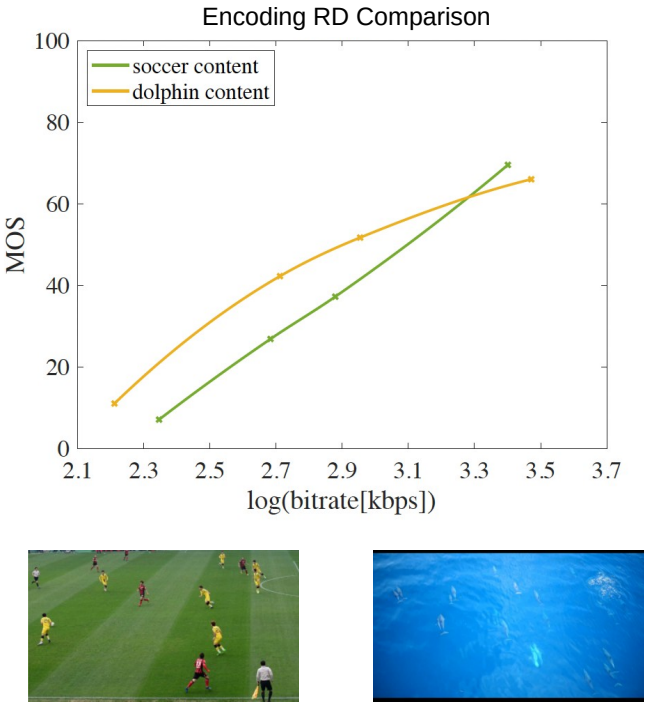


Figure 2.4: Samples of RD curves for different video contents with the same SI and similar TI

the video image so that the filter control analysis is performed to enhance the final decoded video quality. At the end, the encoder uses source coding methods to compress the general control data, the quantized transform coefficients, the intra-prediction data, the inter-prediction data, and the filter control data into a coded bitstream for transmission or storage.

In this thesis, the quantization and rate/quality control are the main focuses.

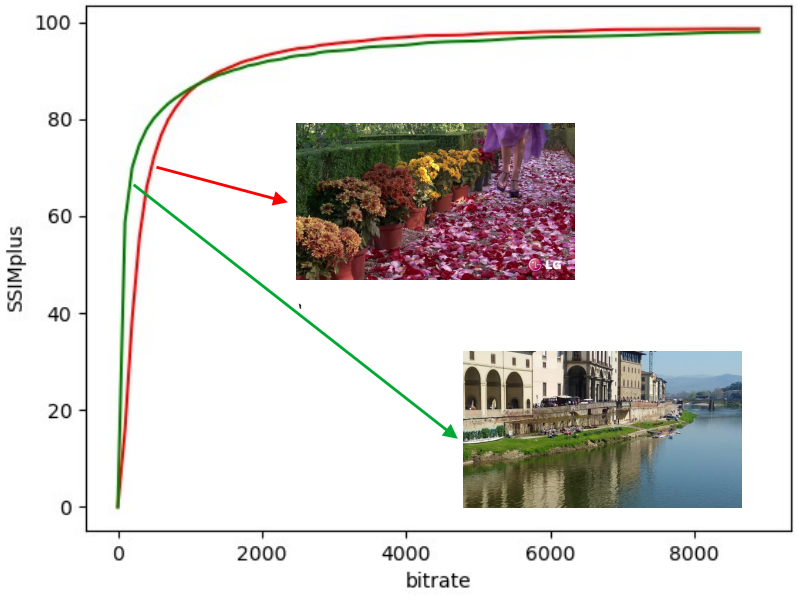
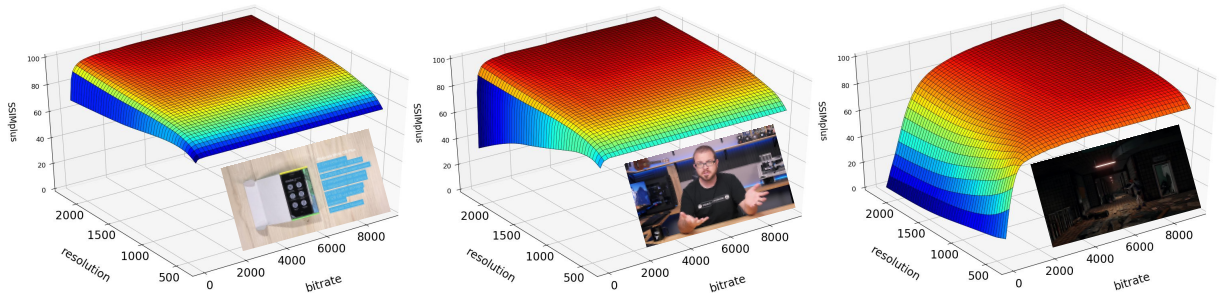


Figure 2.5: Videos with crossing RD curves

2.2.1 Transform and quantization

After all the mode predictions have been finished, no matter how effective a prediction mode selection scheme is used, the final quality and bitrate are determined by the transform and quantization applied to the remaining residual signal. In block-based hybrid coding such as HEVC, transforms are applied to the residuals obtained after the mode prediction. There are two types of transforms used in HEVC: the core transform based on the discrete cosine transform (DCT) and the alternate transform based on the discrete sine transform (DST). After transforming the residual from spatial domain to frequency domain, in order to achieve different levels of compression, a quantization step is necessary.

In signal processing, quantization is a process that maps input from a large set of possible values to an output value in a smaller set. Through mapping multiple values to



(a) (b) (c)

Figure 2.6: Samples of RD surfaces for different video content.

a single value, the higher compression ratio could be achieved using the entropy coding. Scalar quantization is usually done by rounding and truncation. In video encoder, quantization consists of division by a quantization step size ($Qstep$) and subsequent rounding, while inverse quantization consists of multiplication by the quantization step size. Similar to H.264/AVC [40], a Quantization Parameter (QP) is used to determine the quantization step size in HEVC. QP can take 52 values from 0 to 51 for 8-bit depth video sequences. An increase of 1 in QP means an increase of the quantization step size by approximately 12 percent (i.e., $2^{1/6}$). An increase of 6 leads to an increase in the quantization step size by a factor of 2[41]. In addition to specifying the relative difference between the step-sizes of two consecutive QP values, there is also a need to define the absolute step-size associated with the range of QP values. The following equation shows the relationship between quantization step size Q_{step} and QP.

$$Qstep(QP) = (2^{1/6})^{(QP-4)} \tag{2.15}$$

As shown in Fig 2.8, the quantization step size increases non-linearly with respect to QP.

2.2.2 RD analysis based video rate control

RD analysis is an essential step in encoder rate control. For real world applications, the transmission bandwidth and storage space are limited so that the signal is required

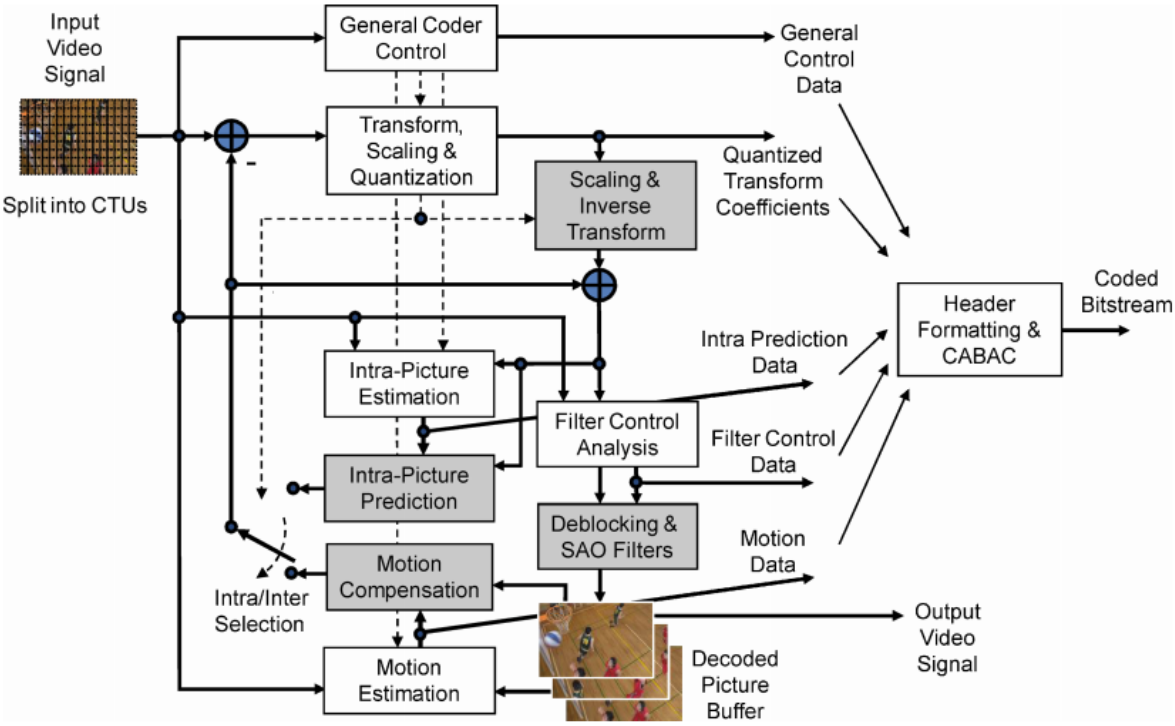


Figure 2.7: HEVC video encoder structure (decoder modeling elements shaded in grey)

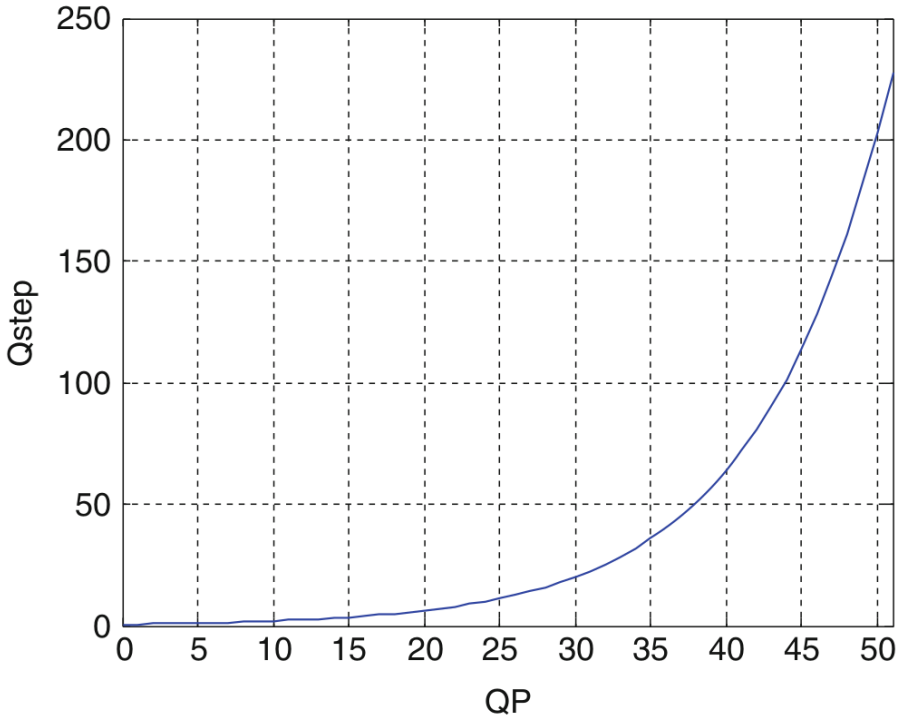


Figure 2.8: Relationship between $Qstep$ and QP

to be transmitted or stored under a predefined data rate. Due to the fact that video signal intrinsically contains a large amount of information both in spatial domain and time domain, it is critical to control its bitrate, which is usually measured in unit of bits/second for video, through compression. The goal of rate control is to keep the bitrate of video below or equal to a certain threshold while selecting the set of coding parameters that achieves the best quality among all possible combinations. The rate control problem can be formulated as finding the set of coding parameters so that the distortion D is minimized, subject to the condition that the encoded video bitrate R is less than or equals to the target bitrate R_t.

In HEVC, the success of rate control is backed by better RD modelling, which guides the modelling between rate and encoder parameters. The RD trade-off can be mathematically formulated:

$$\{\text{Params}\}_{optimal} = \arg \min_{\{\text{Params}\}} D \quad \text{s.t.} \quad R \leq R_t \tag{2.16}$$

where {Params} is the set of coding parameters including intra-prediction modes, inter-prediction modes, and QP etc.. With the help of Lagrange multiplier methods, this hard constrained optimization problem in Eq 2.16 can be relaxed into a soft constraint problem

$$\{\text{Params}\}_{optimal} = \arg \min_{\{\text{Params}\}} D + \lambda R \tag{2.17}$$

where λ is the Lagrange multiplier, which corresponds the slope of the RD curve. The term $D + \lambda R$ is usually treated as the RD cost. In Fig 2.9[42], the best operation point with optimal RD cost for the specific λ can be achieved at the intersection between the line of the cost function $J = D + \lambda R$ and the RD curve.

Based on the idea of RD trade-off, for HEVC and later developed video encoders such as AOMedia Video 1 (AV1), a λ-domain rate control algorithm is proposed to tackle the rate control problem [43].

For the λ-domain rate control algorithm, firstly, an analytic function is used to model the encoded video RD curve.

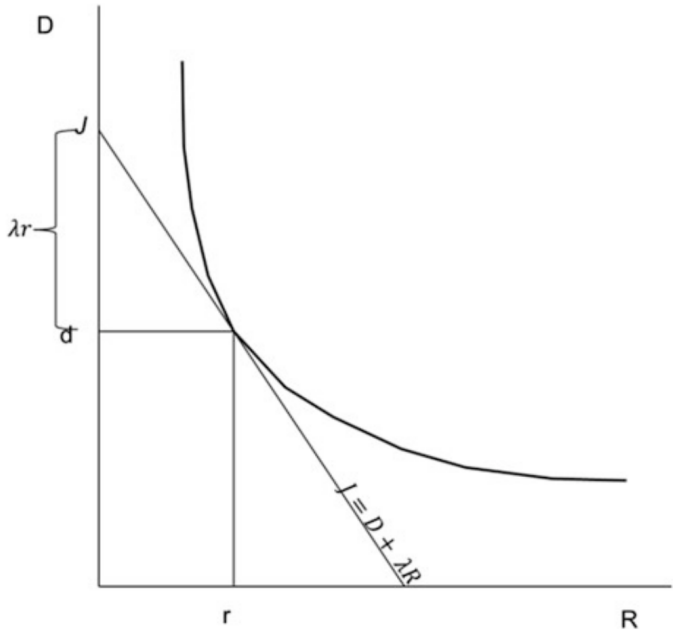


Figure 2.9: Typical RD curve and cost function J with slope $-\lambda$

$$D(R) = CR^{-K} \tag{2.18}$$

where C and K are model parameters related to the characteristic of source video content. The value of R is determined by Eq 2.19.

$$bpp = R/(f \cdot w \cdot h) \tag{2.19}$$

where R is the target bitrate, f is the video frame rate, w is the frame width measured in pixels, and h is the frame height measured in pixels. It should be noted that in practical encoder implementation, bits per pixel (bpp) is used to denote the R value in Eq 2.18.

As mentioned in Fig 2.9, λ is the slope of the RD curve. Therefore, it can be expressed as

$$\lambda = -\frac{\partial D}{\partial R} = CK \cdot R^{-K-1} \triangleq \alpha R^{\beta} = \alpha bpp^{\beta} \tag{2.20}$$

where α and β are parameters related to the source video content. In practical HEVC implementation, the Eq 2.20 is repeatedly calculated for the basic coding unit, i.e., CTU, to determine the value of λ for the Rate-Distortion Optimization (RDO) process.

The author of [43] uses only four video sequences with four compression levels each to demonstrate the effectiveness of RD modelling by the analytic function Eq 2.18, which is shown in Fig 2.10.

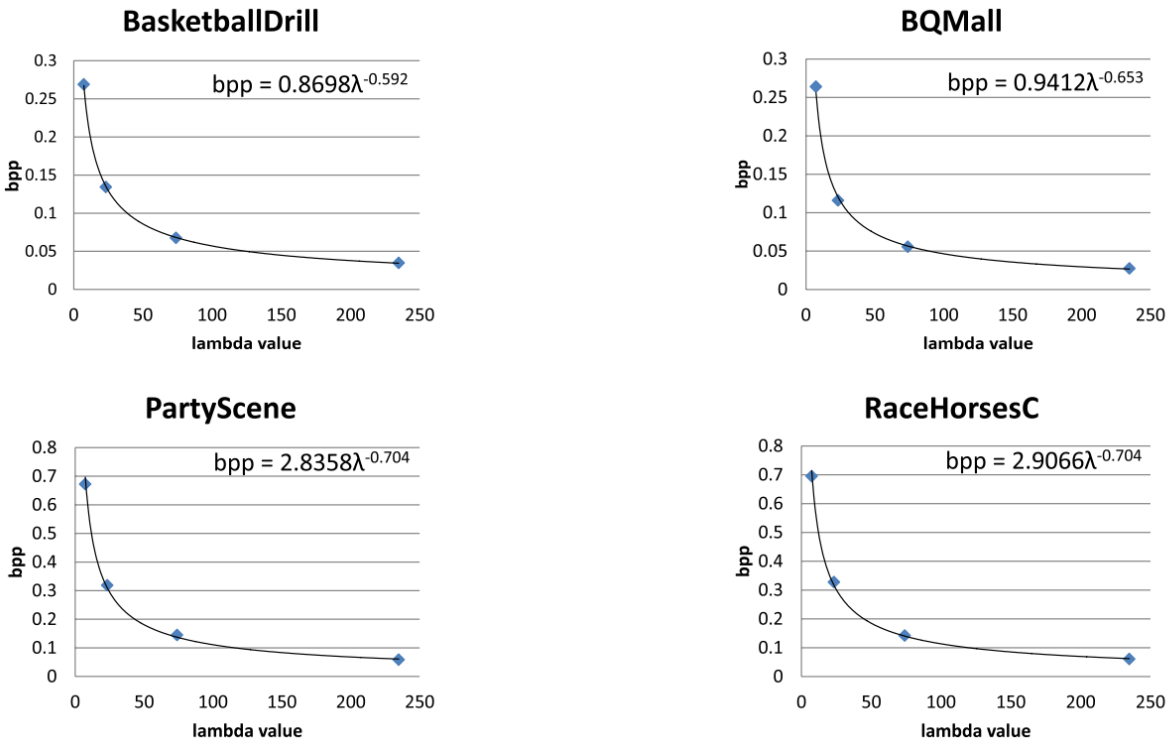


Figure 2.10: RD curve fitting for 4 video sequences

Because the parameters α and β need to adapt to the content of each CTU, an algorithm to update their values is also proposed, as shown below. A more detailed proof can be found at the appendix of [43].

$$\lambda_{comp} = \alpha_{old} bpp_{real}^{\beta_{old}} \tag{2.21}$$

$$\alpha_{new} = \alpha_{old} + \delta_{\alpha} \times (ln\lambda_{real} - ln\lambda_{comp}) \times \alpha_{old} \tag{2.22}$$

$$\beta_{new} = \beta_{old} + \delta_{\beta} \times (ln\lambda_{real} - ln\lambda_{comp}) \times lnbpp_{real} \tag{2.23}$$

In Eq 2.21, Eq 2.22, and Eq 2.23, λ_{comp} is the λ used for the RDO process for the current CTU. α_{old} and β_{old} are the pre-calculated α and β values, respectively, according to the previous round of update. bpp_{real} is the bits per pixel allocated to the current CTU according to the bit allocation process beforehand. λ_{real} is the real λ calculated based on the previous encoded frame. δ_{α} and δ_{β} are the two predefined step sizes. It should be noted that the initial values of α and β are set to 3.2003 and -1.367, respectively, which are given empirically by the authors.

Because QP still needs to be determined for the CTU's encoding, in Eq 2.24, a mapping from λ to QP is defined

$$QP = c_1 \times ln(\lambda) + c_2 \tag{2.24}$$

where c_1 and c_2 are two predefined values set to 4.2005 and 13.7122, respectively. By testing out the algorithm on only 20 video sequences, the author of [43] claims that the aforementioned method can achieve up to 1.81 dB PSNR reduction.

Even though the RD analysis based λ domain rate control enjoys performance gain against previous rate control methods, there are still some shortcomings.

Firstly, the analytic modelling of the RD curve may not be accurate enough. The author claims that the correlation coefficient can achieve 0.99. However, the observation is based on only four source video contents which are encoded at only four distortion levels as shown in Fig 2.10.

Secondly, the update of the parameters of α and β introduces two new heuristic parameters δ_{α} and δ_{β}. Although the authors set the two values to 0.1 and 0.05, respectively, in their experiment, in HEVC implementation the two values, ranging from 0.01 to 0.4 and 0.005 to 0.2 respectively, are set according to the *bpp*, which voids the mathematical derivation of the update rule.

Thirdly, in order to determine the QP value to encode the CTU, the authors link the QP to λ, which in turn links QP to bitrate again according to Eq 2.18. Interestingly, the practice of linking QP with bitrate has already been criticized in the paper for not considering the increased number of coding modes in HEVC.

Though the RD analysis improves the performance of HEVC rate control, the current rate control method still has a large room for improvement. The mathematically derived solutions still need to be guided by heuristic modelling and tuning in real video encoder applications. In real world applications such as x265 video encoder, the heuristic modelling fails frequently. In order to compensate for the short usage of the given bandwidth, redundant bits are appended to the encoded bitstream. And dropping video frames is common if the actual encoded signal rate is higher than the target rate. The inaccurate RD modelling lead to inefficient use of the precious bandwidth in the encoded bitstream.

2.2.3 Video encoder quality control

Quality control is necessary for many research applications such as encoder performance comparison and visual quality database construction. For modern video encoders, users can select the quality level by adjusting the values in constant rate factor (CRF) encoding mode. The mode limits the quality variation across the frames, while rate control and Constant Quantization Parameter (CQP) methods would result in degraded quality in high motion or complex texture part of the video sequence. Moreover, since the CRF encoding mode would disable the rate control mechanisms that lead to lower-than-claimed RD performance, many video quality related datasets make use of CRF encoding mode in database construction.[44, 45, 46, 47]

Most viewers care about the quality of encoded video. Therefore, video quality has always been an active research area and many HVS inspired objective video quality metrics are proposed such as SSIMplus[8] and VMAF[48]. They are becoming widely accepted as the necessary tool to assess the encoded video quality in both academic research and industrial applications.

In the reference encoder of HEVC, constant quality across the frames can only be achieved using CQP setting for all frames. Since each QP value comes with a predefined

quantization matrix, the constant QP mode cannot adapt according to the content, resulting in large quality variations between different frames depending on their complexity. As a solution to the problem, CRF encoding mode mechanism would adapt the QP value according to the content complexity, especially motion information, of each frame[49]. Unlike RD analysis backed rate control mechanism, the CRF method employs a much simplified modelling of the quality against encoding parameter QP.

In real world applications of x264[50] and x265[51], CRF values range from 0 to 51, where 0 represents the lossless quality and 51 represents the worst quality the encoded video can achieve. During the encoding of each frame, the look ahead mechanism would make use of the residual signal to analyze the complexity of the frame ahead to adjust the quantizer curve compression factor, which can be symbolized as $qComp$. Therefore, the $QStep$ in 2.15 can be divided by the $qComp$ as follows:

$$Qstep(QP) = (2^{1/6})^{(QP-4)}/qComp \tag{2.25}$$

where $qComp$ is defined in the range of $(0, 1]$, which increases when the complexity is high so that the quantizer curve in Fig. 2.8 would change accordingly. The described formulation of the CRF control process is a simplified version but captures the essential idea of how modern video encoders achieving constant quality.

There are several drawbacks of the current constant quality control mechanism. The CRF method falls short of the growing expectation of better quality control given the development of HVS inspired quality metrics:

Firstly, adapting the quantization curves merely based on the analysis on residual is not enough. Since the residual signal is obtained after mode prediction as described in section 2.2.1, the values are greatly affected by the inter or intra prediction mode selected, which is far from an adequate prediction of frame complexity.

Secondly, the actual relationship between content characteristics and QP is unknown and has never been thoroughly investigated in the research area. The current widely used quality control methods are mostly based on heuristics and ad-hoc mechanisms such as using the quantizer curve compression factor $qComp$ described above.

Thirdly, the established HVS inspired metrics such as SSIMplus and VMAF has been validated as meaningful metrics that well capture the quality of encoded videos effectively, much better than the signal error estimators such as Peak Signal to Noise Ratio (PSNR) and Mean Square Error (MSE). Therefore, adjusting encoded video quality according to HVS inspired metrics is highly desirable and beneficial to many tasks such as video database construction or video encoder comparison.

Both the rate and quality control of video encoders boil down to the problem of modelling encoder control parameters against the target rate and the quality of the encoded videos. With the help of the RD analysis conducted in the rate control mechanism of HEVC, a better quality control could not only make the encoded video achieve less quality variation but also enable the encoded video to achieve the target quality level measured by HVS inspired metrics. Without a good estimation of video characteristics, it would be a hard task to find an appropriate set of parameters that are suitable for different content. Therefore, a thorough investigation and better mathematical modelling on visual content characterization from the RD perspective would benefit the encoder quality control. The work in Chapter 4 tackles the problem from an RD analysis perspective and solves it by introducing a better quality-parameter modelling method.

2.3 End-to-End Image Compression

In 2017, Ballé et al. proposed a neural network based E2E image compression framework.[12] The encoder and decoder are comprised of convolutional neural networks followed by non-linear functions. By using RD as the cost function, the proposed work achieves much better performance when compared against conventional image encoders such as JPEG and JPEG 2000, especially measured in HVS inspired metrics such as MS-SSIM[52]. Following the paper, many works have been done based on the idea of E2E optimized RD balance convolutional framework [53, 54, 55]. In the work proposed by Minnen et al [53], the authors make use of the Gaussian hyperprior assumption of the encoded parameters, achieving better performance against HEVC intra coding, the best performing transform-based image encoder.

2.3.1 General framework

The goal of signal compression is to achieve the best quality under limited data rate budget, which can be expressed mathematically as an RDO problem. The RD cost function is given by $R + \lambda D$, where R represents the data rate, usually measured in the unit of bits per second or bits per sample, D represents the distortion, usually measured in mean square error or perceptually inspired quality metrics, and the Lagrange multiplier λ balances the trade-off between rate and distortion. The proposed general nonlinear transform coding framework directly optimizes for the RD cost function in an E2E manner.

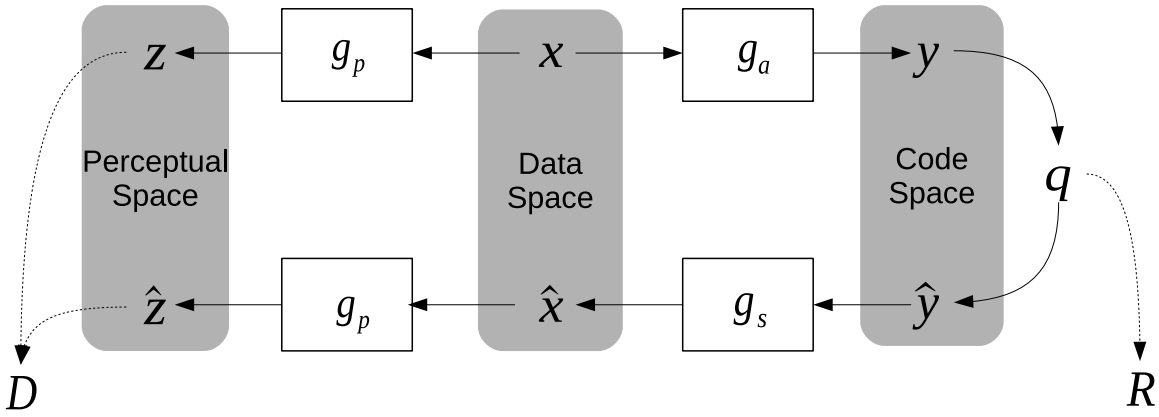


Figure 2.11: Autoencoder in a transform coding framework.

Fig.2.11[12] illustrates the coding framework, where x and $\hat{x}$ represent the original pristine and reconstructed signals in data space, respectively, y and $\hat{y}$ represent the corresponding coded signals in the continuous code space, and z and $\hat{z}$ represent the transformed signals in the perceptual space for quality purpose. In this framework, a vector of signal values $x \in \mathbb{R}^N$ is mapped to the latent code space by a parametric analysis transform, $y = g_a(x, \phi)$, where ϕ represents the vector of parameters that need to be optimized. After the analysis transform, the coding space representation y is quantized, producing a discrete-valued vector $q \in \mathbb{Z}^m$, which is compressed afterwards using context-adaptive binary arithmetic coding (CABAC)[56]. The rate R of the discrete code is lower-bounded by

the entropy of the quantized vector $H[P_q]$. To reconstruct the signal, the quantized signal is mapped back to the continuous code space. Then the parametric synthesis transform is applied on the dequantized signal $\hat{y}$, $\hat{x} = g_s(\hat{y}, \theta)$, where θ represents another vector of parameters that need to be optimized. The distortion is computed by transforming to a perceptual space using the transform $\hat{z} = g_p(\hat{x})$, followed by evaluating a distortion metric $d(z, \hat{z})$. The perceptual transform in the proposed method is the identity transform, but other perceptually meaningful transforms can be applied as well. The parameter vectors ϕ and θ are optimized for a weighted sum of the rate and distortion measures, $R + \lambda D$, over a set of images. As in [12], CNNs are used to implement the analysis and synthesis transforms, allowing for end-to-end learning and testing. There are three layers in the analysis transform, each consisting of a convolutional layer (with 128 kernels of sizes 9, 5, 5), followed by a down-sample layer (of downsampling factors 4, 2, 2) and a Generalized Divisive Normalization (GDN) layer. The synthesis transform is the inverse of the analysis transform.

The optimization process aims to minimize the RD cost over the parameters of forward, inverse and perceptual transforms. The Lagrange multiplier λ is set to govern the trade-off between rate and distortion. A key difference between the proposed method and conventional image encoding methods is to directly applying RD analysis in an end-to-end manner. Furthermore, the nonlinear transform is used to warp the space appropriately instead of searching for the optimal quantization scheme over the high dimensional signal space which is nearly intractable. The warping process makes it possible to use a fixed uniform scalar quantizer in code space, and largely simplifies the coding process. The objective function is defined in terms of entropy as

$$L[g_a, g_s, P_q] = -\mathbb{E}[\log_2 P_q] + \lambda \mathbb{E}[d(z, \hat{z})] \tag{2.26}$$

where P_q is the probability mass function of the quantized output vector of the analysis transform.

A technical difficulty is that the derivatives of the quantization function are zero almost everywhere, making it impossible to execute any gradient descent based optimization methods. As in [12], the quantizer is replaced with an additive i.i.d uniform noise source $\triangle y$, which has the same width as the quantization bins (one). Consequently, the continuous

relaxation density function of $\tilde{y} = y + \triangle y$ can be used in the gradient descent process

$$p_{\tilde{y}}(n) = P_q(n), \text{ for all } n \in \mathbb{Z}^M \tag{2.27}$$

With the continuous approximation of the quantized coefficient distribution, the loss function for parameters θ and ϕ across all training samples i is

$$\begin{aligned} L(\theta, \phi) = \mathbb{E}_{x, \triangle y}[- \sum_i \log_2 p_{\tilde{y}_i}(g_a(x; \phi) + \triangle y; \psi^{(i)}) \\ + \lambda d(g_p(g_s(g_a(x; \phi)) + \triangle y; \theta), g_p(x))] \end{aligned} \tag{2.28}$$

Mean squared error (MSE) is chosen as the distortion measure d, though any other differentiable quality metric can be adopted in the general framework.

2.3.2 Performance

In terms of performance, the proposed framework greatly outperforms conventional transform based encoders such as JPEG and JPEG2000. Fig.2.12 shows an aggregate performance comparison of the three encoders on Kodak dataset. Since quality and rate cannot be fixed across different images and encoders, the points on the average curves are the mean RD points connected by the 24 encoded source content images' RD data. Two different modes, target quality and target rate, are tested for JPEG 2000 since the two modes lead to different image encoder parameter selections.

It can be seen from the Fig.2.12 that the proposed work achieves the best aggregated performance for all RD levels. Fig.2.13, Fig.2.14 and Fig.2.15 are three examples of the content-wise performance comparison including a visual comparison of encoded images.

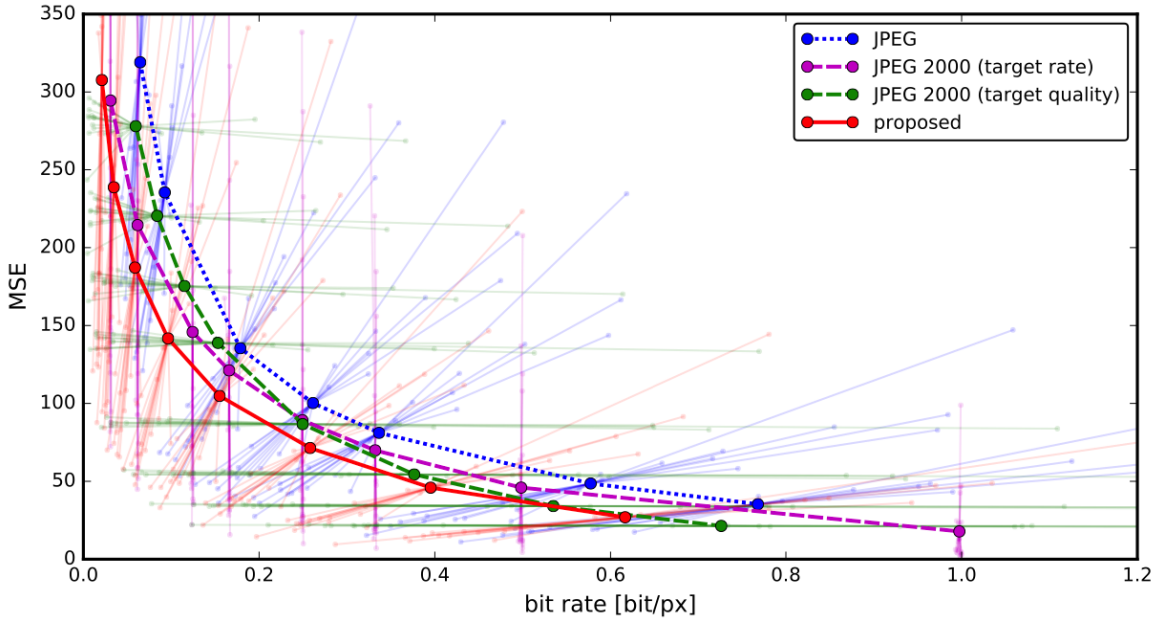
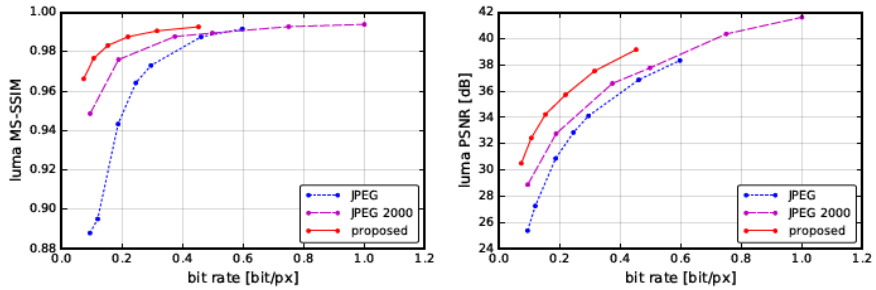


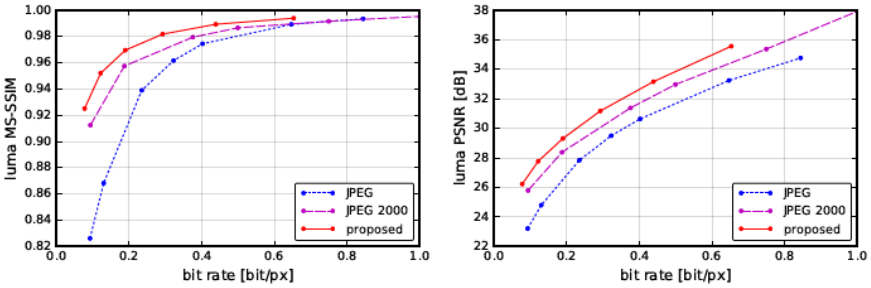
Figure 2.12: Summary rate-distortion curves.



Proposed method, 3749 bytes (0.106 bit/px), PSNR: luma 32.43 dB/chroma 34.00 dB, MS-SSIM: 0.9767

JPEG 2000, 3769 bytes (0.107 bit/px), PSNR: luma 29.49 dB/chroma 32.99 dB, MS-SSIM: 0.9520

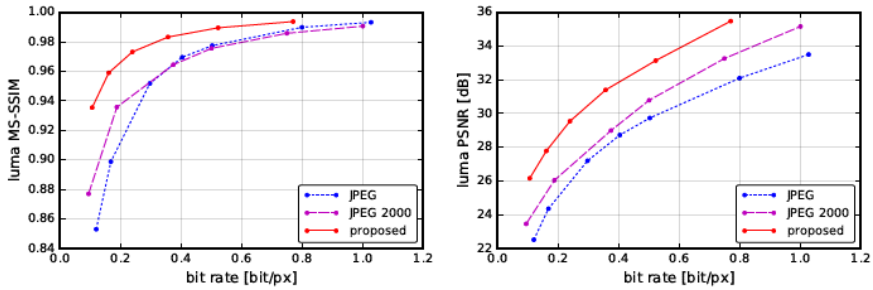
Figure 2.13: Natural Scene Example



Proposed method, 6680 bytes (0.189 bit/px), PSNR: luma 29.31 dB/chroma 36.17 dB, MS-SSIM: 0.9695

JPEG 2000, 6691 bytes (0.189 bit/px), PSNR: luma 28.45 dB/chroma 35.32 dB, MS-SSIM: 0.9586

Figure 2.14: City View Example



Proposed method, 5683 bytes (0.161 bit/px), PSNR: luma 27.78 dB/chroma 32.60 dB, MS-SSIM: 0.9590

JPEG 2000, 5724 bytes (0.162 bit/px), PSNR: luma 25.36 dB/chroma 31.20 dB, MS-SSIM: 0.9202

Figure 2.15: Drawing Example

2.3.3 Observation and discussion

Though the E2E encoding framework outperforms conventional image encoders by a large margin, it can hardly be applied in real world applications due to the lack of an accurate control mechanism to guide the image encoder to select the appropriate λ value for a specific quality or rate level. Without knowledge of λ-quality or λ-rate relationship, quality or rate control for encoded images is not feasible since the source image complexity varies across different content. As shown in Fig.2.13, Fig.2.14 and Fig.2.15, rate-distortion behavior varies substantially across bit rates for different contents. Image space contains tens of thousands of contents. Each content has its own characteristics, leading to drastically different RD behaviors. It would be beneficial to analyze the characteristics of image in terms of RD so that encoding parameters can be chosen appropriately for the target quality level or data rate.

In order to conduct a controlled ablation experiment for deeper analysis, only one parameter should be changed while all other parameters are fixed to the same level. For encoder performance comparisons such as the one shown in Fig.2.12, the rate or distortion cannot be aligned across encoders for the same source content compressed by different encoders. This is due to the lack of an accurate and generalized encoding parameter model against quality or rate without sacrificing encoding efficiency. There are two ways to compare encoders' performance. The first is to average RD data points and then interpolate the mean data points. However, since the encoding control methods are not the same across encoders, the average RD behaviours are different. As can be seen from the performance comparison in Fig.2.12, the average data points for each RD curve have to be the average of the 24 individual contents' RD points. For the E2E method the points are grouped and averaged based on the same λ value. For JPEG 2000, the target quality points are averaged according to the same quality level while in target rate setting points are averaged for the same rate level. The averaging target choice has a substantial effect on the RD plots since different choices lead to the average over different set of RD points. The other way to compare encoders' performance is to obtain the RD curves for each individual content across different encoders first. Using the Bjøntegaard Delta (BD) method [57][58] by interpolating the sparse points with polynomial functions, one can then calculate the

rate saving by integration. However, as discussed by Duanmu et al.[1], the BD-rate method is prone to produce non-monotonic RD curves and cannot guarantee the inversibility of rate-distortion and distortion-rate curves for the same set of data points. Therefore, the BD-rate method cannot be counted as a reliable way to compare encoder performance. The aligned quality or rate for each compressed image or video is necessary for an objective and accurate evaluation of the encoders performance.

Without the knowledge of λ-quality or λ-rate relationship, which are based on the visual content characteristics such as RD information, the aforementioned E2E image compression control is not possible. In summary, just as its importance in video quality and rate control, image content characterization plays an essential role in E2E neural network driven encoder control as well.

2.4 Summary and Discussion

This chapter begins with the introduction of the low level vision traits for visual content characterization and their usage in both image and video quality databases. Then we introduce visual complexity, which is a concept often used in the area of image compression in evaluating the source contents' compressibility. By classifying the visual complexity traits into filter based measures and compression based measures, we describe the frequently used SI and TI and several Komogrov complexity inspired measures. We conclude that the over-simplified single dimension visual traits can only capture limited perspectives of visual contents. We show that the RD behaviors vary drastically for different visual content with similar SI and TI values. Therefore, there is a strong desire to develop visual content characterization method for image compression and related image quality assessment applications.

We introduce video encoding pipeline focusing on encoder rate and quality control using HEVC as the example. By describing the control mechanism and the complex workflow of video encoder, the necessity of a precise content adaptive control of the encoding parameter becomes manifest. Furthermore, the framework of neural network based E2E image compression is introduced. We apply the RD analysis inspired quality control model in

Chapter 4. We will also use it for testing the source content selection method in Chapter 5.

Chapter 3

Encoder Performance Analysis and Observations

3840 × 2160 or 4096 × 2160 pixel resolution (4K), Ultra High Definition (UHD), and higher resolution video contents have become increasingly popular recently. The largely increased data rate casts great challenges to video compression and communication technologies. Emerging video coding methods are claimed to achieve superior performance for high-resolution video content, but thorough and independent validations are lacking. In this study, we carry out an independent and so far the most comprehensive subjective testing and performance evaluation on videos of diverse resolutions, bit rates and content variations, and compressed by popular and emerging video coding methods including H.264/Advanced Video Coding (AVC), H.265/High Efficiency Video Coding (HEVC), VP9, Audio Video Coding Standard 2 (AVS2) and AOMedia Video 1 (AV1). Our statistical analysis derived from a total of more than 36,000 raw subjective ratings on 1,200 test videos suggests that significant improvement in terms of Rate-Distortion (RD) performance against the AVC encoder has been achieved by state-of-the-art encoders, and such improvement is increasingly manifest with the increase of resolution. Most importantly, based on the database construction process and encoder performance comparison through RD analysis, the problems of video quality parameter modelling for precise encoded video quality control, limitations of widely used visual features for compression datasets, and

source content selection are discovered, revealed, and discussed.

3.1 Introduction

4K, UHD, and higher resolution video contents have enjoyed a remarkable growth in recent years. 4K/UHD (4096×2160 or 3840×2160) video increases the resolution by a factor of four from Full High Definition (FHD) (1920×1080) and offers significantly increased sharpness and fine details. 4K/UHD video displays are believed to deliver better Quality of Experience (QoE) to viewers and are becoming widely available on the consumer market.

While 4K/UHD videos raise the potentials for better user QoE, their higher data rates cast great challenges to video distributions, for which video compression technologies are crucial in controlling the bandwidth of video so as to fit the distribution pipeline. The currently most widely used video coding technologies based on H.264 AVC standards hardly meet the requirement. To this end, several modern video encoders including H.265 HEVC [59], AV1 [60], and AVS2 [61] are deliberately optimized for compressing content of 4K and higher resolutions. With many video encoders at hand, it becomes pivotal to compare their performance, so as to choose the best algorithms and find the direction for further advancement. Because the Human Visual System (HVS) is the ultimate receiver in most applications, subjective evaluation is a straightforward and reliable approach to evaluate the quality of videos. Although expensive and time consuming [62], a comprehensive subjective study has several benefits. First, it provides useful data to study human behaviors in evaluating perceived quality of encoded videos. Second, it supplies a test set to evaluate and compare the relative performance of classical and modern video encoding algorithms. Third, it is useful to validate and compare the performance of existing objective video quality assessment (VQA) models in predicting the perceptual quality of encoded videos. This will in turn provide insights on potential ways to improve them.

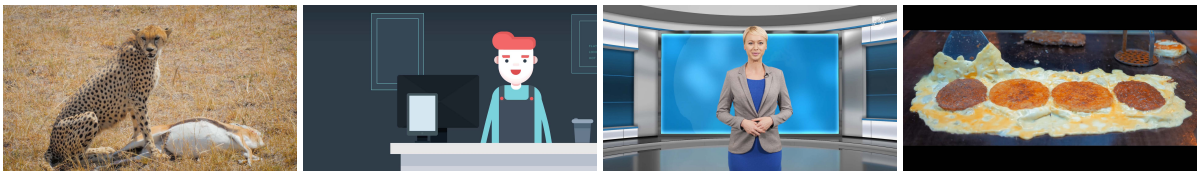
Several recent subjective studies have been conducted to evaluate the encoder performance on 4K video compression [63, 64, 65, 66]. It is generally observed that the latest video encoders can deliver 4K contents with better viewer QoE, although the test only covers a small number of contents. In addition, most of the work covers FHD and 4K for

HEVC and AVC encoders only. In [67], HEVC encoder is evaluated by using 10 contents under 4K resolution. In [68], the performance of HEVC, AVC, and VP9 [69] at FHD and 4K are compared on 10 contents, from which it is shown that HEVC and VP9 achieve better bitrate reduction than AVC at the same quality level. The performance of the emerging next-generation encoders, AV1 and AVS2, on 4K videos has not been systematically evaluated. In summary, all of the aforementioned studies suffer from the following problems: (1) the test dataset is limited in size; (2) the types of encoders do not fully reflect the state-of-the-art; and (3) the spatial resolutions do not cover commonly used display sizes. Moreover, many tests have been conducted by the developers or participants of the coding standards. Independent datasets and test results commonly available to the public is lacking.

In this work, we conduct subjective evaluation of popular and emerging video encoders on 4K content. Our contributions are twofold. First, we carry out an independent and so far the most comprehensive subjective experiment to evaluate the performance of modern video encoders including AVC [50], VP9 [69], AV1 [70], AVS2 [71] and HEVC [72]. Second, we applied statistical analysis on the subjective data and observe some significant trends.

3.2 Video Database Construction and Subjective Experiment

The video database is created from 20 pristine high-quality videos of UHD resolution (3840×2160, progressive) selected to cover diverse content types, including humans, plants, natural scenes, architectures and computer-synthesized sceneries. All videos have the length of 10 seconds [73]. The detailed specifications are listed in Table 3.1 and the screenshots are shown in Fig. 3.1. Spatial Information (SI) and Temporal Information (TI) [4] that roughly reflect the complexity of the video content are also given in Table 3.1, which suggests that the video sequences are of diverse spatio-temporal complexity and widely span the SI-TI space. Using the aforementioned video sequences as the source, each video is encoded with AVC, VP9, AV1, AVS2 and HEVC encoders with progressive scan at three spatial resolutions (3840×2160, 1920×1080, and 960×540) and four distortion levels. The



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o) (p)

(q) (r) (s) (t)

Figure 3.1: Snapshots of source video sequences. (a) Safari. (b) 2D cartoon. (c) News. (d) Teppanyaki. (e) Screen recording. (f) Botanical garden. (g) Tears of steel. (h) Soccer game. (i) Animation. (j) Motor racing. (k) Climbing. (l) Colorfulness. (m) Forest. (n) Lightrail. (o) Dolphins. (p) Dance. (q) Spaceman. (r) Barbecue. (s) Supercar. (t) Traffic.

detailed encoding configurations are as follows:

- HEVC: We employ x265 [72] with main profile for HEVC encoding. The Group of Pictures (GOP) size is set to 60. Rate control mode is selected to be constant rate factor (CRF). Videos are encoded in "veryslow" speed setting.
- AVC: The x264 [50] with high profile of level 5 is used for AVC encoding. Other settings such as GOP size, rate control mode and speed setting are the same as those of the HEVC configurations.
- VP9: The libvpx software [69] is used for VP9 encoding. The encoding parameters, such as GOP size, rate control mode, etc., are set to be as similar as possible to HEVC. The parameter selection is based on [74].
- AV1: The AV1 reference software aomenc [70] is used for AV1 encoding. The encoding parameters are set to be as similar as possible to HEVC. The parameter selection is based on [74].
- AVS2: The libxavs2 [71] is used for AVS2 encoding. The encoding parameters, such as GOP size and speed setting are set to be as similar as possible to HEVC. The parameter selection is based on the configuration file "encoder_ra.cfg" that comes with AVS2 source code [71].

A small-scale internal subjective test is conducted and the encoding bitrates are adjusted to ensure that the neighboring distortion levels are perceptually distinguishable. Eventually, we obtain 1,200 videos encoded by 5 encoders in 3 resolutions at 4 distortion levels.

Our subjective experiment generally follows the single stimulus methodology as suggested by the ITU Telecommunication Standardization Sector (ITU-T) recommendation P.910 [4]. The experiment setup is normal indoor home settings with ordinary illumination level and no reflecting ceiling walls or floors. All videos are displayed at 3840×2160 resolution on a 28 inch 4K LED monitor with Truecolor (32bit) at 60Hz. The monitor is calibrated to meet the ITU-T BT.500 recommendations [75]. Videos are displayed in

Table 3.1: SI, TI, Frames per Second (FPS), and Description of Source Videos

Name	FPS	SI	TI	Description
Safari	24	26	41	Animal, smooth motion
2D carton	25	38	55	Animation, camera motion
News	25	32	45	Human, static
Teppanyaki	24	33	32	Food, average motion
Screen recording	30	82	12	Screen content, partial motion
Botanical garden	30	112	10	Natural scene, static
Tears of steel	24	28	61	Movie, high motion
Soccer game	30	54	24	Sports, high motion
Animation	30	55	32	Animation, high motion
Motor racing	24	57	37	Sports, camera motion
Climbing	30	38	73	Game, high motion
Colorfulness	30	23	65	Texture, smooth motion
Forest	24	46	24	Natural scene, camera motion
Lightrail	30	79	32	Architecture, camera motion
Dolphins	25	54	23	Animal, smooth motion
Dance	30	73	32	Human, high motion
Spaceman	24	51	2	Human, static
Barbecue	25	100	11	Natural scene, smooth motion
Supercar	25	80	22	Sports, average motion
Traffic	30	89	24	Architecture, high motion

random order using a customized graphical user interface from which individual subjects' opinion scores are recorded.

A total of 66 naïve subjects, including thirty nine males and twenty seven females aged between 18 and 35, participated in the subjective test. Visual acuity and color vision are confirmed with each subject before the subjective test. To familiarize the subjects with the testing environment, a training session is performed before the formal experiment, in which 3 videos different from those in the formal experiment are rendered. The same methods are used to generate the videos used in the training and testing sessions. Therefore, before the testing session, subjects knew what distortion types would be expected. Subjects were instructed with sample videos to judge the overall video quality based on the distortion level. Due to the limited subjective experiment capacity, we employed the following strategy. Each subject is assigned 10 contents in a circular fashion. Specifically, if subject i is assigned contents 1 to 10, then subject $i+1$ watch contents 2 to 11. Each video is assessed for at least 30 times and more than 36,000 subjective ratings are collected in total. For each subject, the whole study takes about 3 hours, which is divided into 6 sessions with five 5-minute breaks in-between to minimize the influence of fatigue effect.

We employ 100-point continuous scale as opposed to a discrete 5-point ITU Radiocommunication Sector (ITU-R) Absolute Category Scale (ACR) for three advantages: broader range, finer distinctions between ratings, and demonstrated prior efficacy [76]. After converting the subjective scores to Z-scores per session to account for any differences in the use of the quality scale between sessions, we proceed to an outlier removal process suggested in [75]. No outlier detection is conducted participant-wise. After outlier removal, Z-scores are linearly re-scaled to lie in the range of [0, 100]. The final quality score for each individual video is computed as the average of the re-scaled Z-scores, namely the mean opinion score (MOS), from all valid subjects. Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SRCC) between the score given by each subject and MOS are calculated. The average PLCC and SRCC across all subjects are 0.79 and 0.78, with standard deviation (STD) of 0.09 and 0.08, respectively, suggesting that there is considerable agreement among different subjects on the perceived quality of the test video sequences.

3.3 Encoder Performance Analysis

We use the MOS of the test videos described in the previous section to evaluate and compare the performance of the encoders. It is worth noting that the performance comparison is based on the encoder configuration provided earlier, where all encoders are set to configurations equivalent to the 'veryslow' setting of the HEVC encoders.

Sample RD curves for individual test videos are given in Fig. 3.2. From the RD curves of all content, we have three observations.

First, AVC under-performs all the other four encoders in most cases, which can be justified by the increased flexibility in almost every prediction coding part of recent encoders. For instance, during the process of intra-prediction, AVC only supports intra mode in 9 directions. Comparatively, the number of directions becomes 10 for VP9, 33 for AVS2, 35 for HEVC, and 56 for AV1. In addition to increasing number of prediction directions, AVS2, VP9 and AV1 add another level of flexibility by increasing number of possible partition modes. In AVC and HEVC, the current Coding Unit (CU) can only be partitioned into 4 smaller size CUs, if the encoder decide not to further split the CU, the current CU will be directly used for direction prediction[77]. However, in VP9 if the encoder decide there is no need to further split the CU, it still has two more partition options to choose, namely horizontal split and vertical split[78]. The same number of options becomes 4 and 9 for AVS2 and AV1 respectively[79, 78].

Second, the performance difference between different encoders, exhibited as the gaps between the RD curves, become increasingly manifest with the increase of resolution from 540p to 1080p, and then to 2160p. This validates the coding gain obtained by the advanced technologies specifically designed for high resolution videos in the newly developed encoders, for which the increased coding unit size makes the major contribution. In contrast to AVC's basic coding unit size of 16×16, the largest coding unit size becomes 64×64 for HEVC, VP9 and AVS2. For AV1, the size can even go up to as large as 128×128. Larger block size is more efficient in rate distortion sense[80, 81, 82].

Third, we observe that different video content lead to different RD behaviour. For example in Fig.3.2, the high motion content Tears of steel has an upward trended RD

curves when compared with the Barbecue content at high quality range. The reason may be that the smoke content in the Barbecue video contains a lot of texture that cannot be well predicted using predictive encoding methods.

Table 3.2: Column BD-Rate Saving vs. Row (negative percentages suggest column encoder savings against row)

540p	AVC	HEVC	AVS2	VP9	AV1
AVC	0	-	-	-	-
HEVC	-22.7%	0	-	-	-
AVS2	-20.3%	-4.7%	0	-	-
VP9	-28.9%	-20.5%	-25.7%	0	-
AV1	-34.4%	-23.3%	-17.6%	-4.5%	0

1080p	AVC	HEVC	AVS2	VP9	AV1
AVC	0	-	-	-	-
HEVC	-42.2%	0	-	-	-
AVS2	-45.8%	-9.8%	0	-	-
VP9	-47.5%	-18.5%	-18.1%	0	-
AV1	-48.7%	-20.1%	-21.4%	-3.5%	0

2160p	AVC	HEVC	AVS2	VP9	AV1
AVC	0	-	-	-	-
HEVC	-61.2%	0	-	-	-
AVS2	-63.5%	-9.7%	0	-	-
VP9	-62.2%	-8.7%	-5.3%	0	-
AV1	-63.2%	-9.5%	-15.0%	-16.4%	0

In addition to the qualitative analysis, we also compute the average bitrate saving [57, 58] of each encoder over another. The result is shown in Table 3.2, from which we can observe that on average AV1 outperforms the other encoders with a sizable margin. However, it is worth noting that the RD performance gain by AV1 is highly content dependent

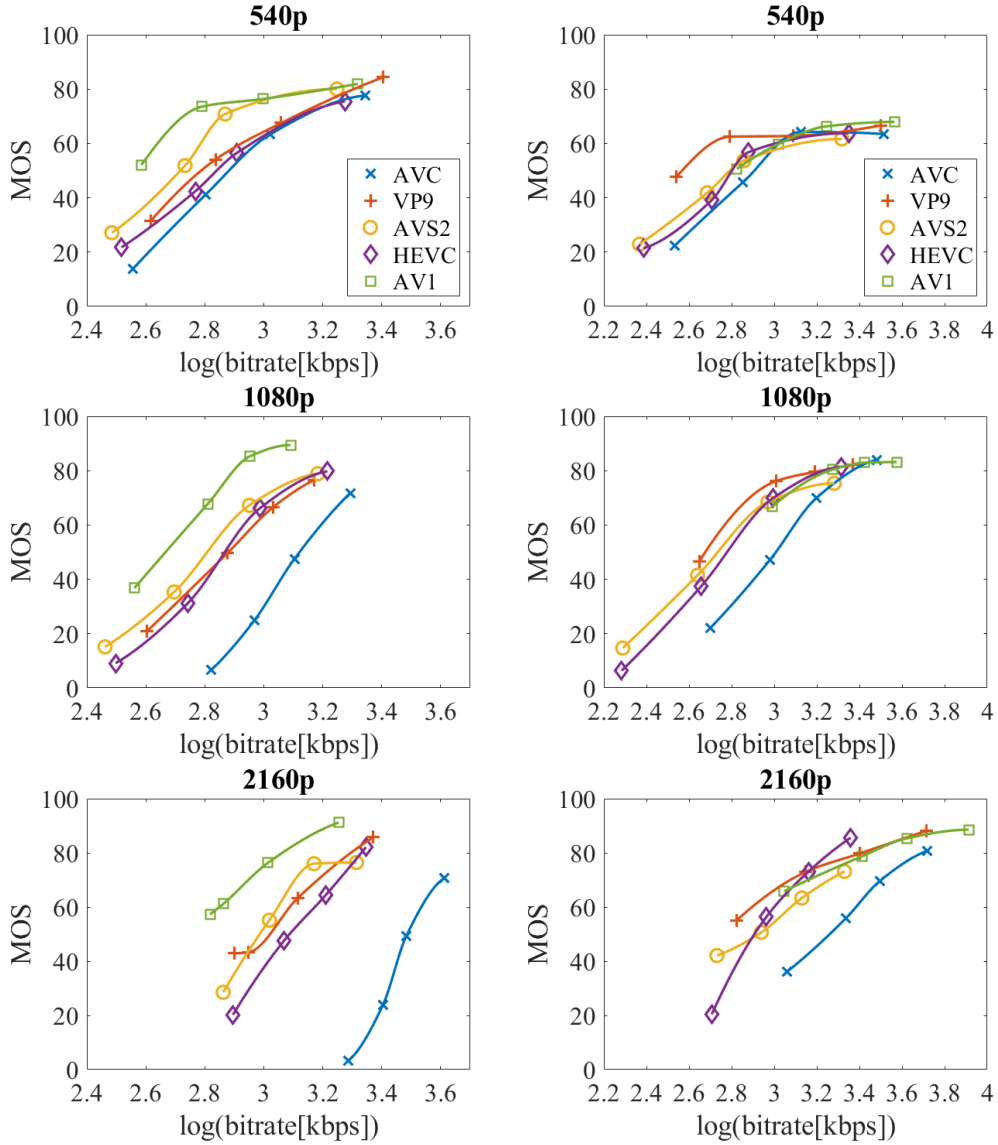


Figure 3.2: RD curves of AVC, VP9, HEVC, AVS2 and AV1 encoders for 540p, 1080p and 2160p resolutions for Tears of steel (left) and Barbecue (right).

and that AV1's performance is achieved on the condition of its much higher complexity compared with all other encoders.

The time complexity performance test is done on a Ubuntu 16.04 system with Intel E5-1620 CPU, 32GB Corsair 2666MHz RAM and Crucial 2TB 530/510 MB/s read/write speed hard drive. As shown in Table 3.3, we can see that AV1 consumes over 500 times of AVC's computational time, which takes the least amount of encoding time. The results suggest that state-of-the-art AVC implementations are still highly competitive choices for time critical tasks, while the encoding speed of AV1 may hinder it from many practical applications. It is worth mentioning that AV1 is still under development and the current version has not been fully optimized for multi-thread encoding. VP9 and HEVC show comparable time complexity, while AVS2 doubles their encoding time. They compromise between compression performance and speed. We have observed that each generation of video encoder outperforms the other encoder in RD sense by a sizable margin. On the other hand, it is almost certain that each generation of video encoders are slower than their previous generation due to that the increased encoding flexibility brings more decisions to be made during encoder mode prediction process. Even though it has always been claimed that video encoder can be sped up by code optimization techniques such as multi-threading, the RD performance has to be sacrificed in exchange for lower time complexity [83, 84].

Table 3.3: Encoder Relative Complexity vs. AVC at 3 Resolutions

	AVC	HEVC	AV1	VP9	AVS2
2160p	1	4.2810	590.74	5.2856	9.8568
1080P	1	4.7314	546.19	6.6286	10.0401
540P	1	5.2805	806.15	5.2572	11.7716

3.4 Observations and Discussion

Besides the encoder performance analysis, which is the original objective of this work, we have several observations regarding the current video encoding framework and the general practice of compression quality database construction:

1. The current quality control CRF method implemented in video encoders is far from perfect. Though the purpose of CRF method is to achieve constant quality levels of the encoded video across video frames by tuning the CRF parameter to achieve different levels of quality, the actual final quality varies significantly and is difficult to predict due to different video contents' distinctive characteristics. As discussed in the background chapter, the constant quality controller implemented in all video encoders utilize variations of visual complexity metric SI for each frame while we have already known that SI is not a reliable metric given the complex RD behaviours of different source contents. During the construction of the dataset, in order to select the encoder CRF that guarantees meaningful separated distortion levels in terms of human perception, we encode source pristine videos into over 10,000 different encoded videos with a range of CRF levels. Then it takes us a significant amount of time to visually check them in order to pick four suitable CRF values that are suitable for all 20 source contents.

Since recent video quality metrics have achieved high correlation against human observers' MOS, it would be much easier to select encoding levels directly using the objective quality metric such as SSIMplus with a set of pre-defined scores (e.g., four scores 90, 75, 60 and 54, representing a wide span of quality levels in practically useful range) as the target quality. Moreover, with a predefined encoded video quality level, average encoder RD performance comparison would be easier. Since videos are encoded into four quality levels that are not aligned across different contents and for different encoders, the interpolated RD curves should be taken with grain of salt. For example in the 2160p tears of steel performance comparison of Fig.3.2, the AV1's RD curve has little overlap with the AVC's RD curve, it is only possible to interpolate the two curves for BD rate comparison. As discussed in [1], the choice of interpolation function has an unpredictable result on the trend of the curves. Therefore, it would be beneficial to have the target quality mode in modern video encoders, where the target quality is measured by objective quality metrics.

2. From the RD performance analysis in this study, we further validate that widely used complexity measures SI and TI are not reliable for source content characterization for compression applications. In the Fig.3.3, we select two contents soccer and dolphin as examples. The two contents have the same SI value of 54 and close TI values of 24 and 23, respectively, while their RD behaviours are different across all the five encoders we

tested. For encoders AV1, VP9, HEVC, and H264, the two contents' RD curves all cross each other. Even though for encoder AVS2, the RD curves do not cross each other, the quality indicated by MOS of dolphin content saturates at around the bitrate of 2.9 kbps while the soccer content has a much worse quality. Therefore, the two single dimensional space of (SI, TI) is extremely limited in describing the complex visual contents containing billions of pixels. Higher dimensional encoding RD analysis is a desirable and potentially more reliable measure for compression applications.

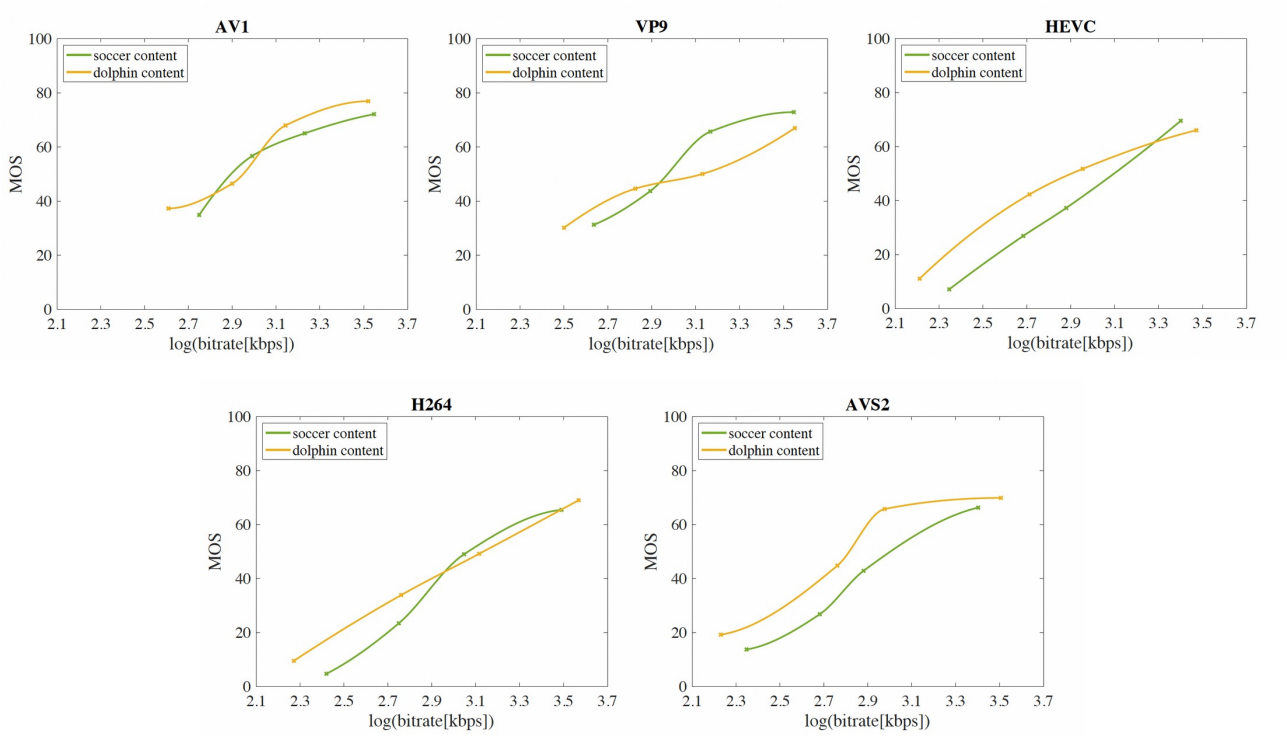


Figure 3.3: RD behaviors for the soccer and dolphin contents across the five encoders

3. The choice of source contents have a great influence on the final encoder performance comparison. As discussed in the RD performance comparison, different video content lead

to different RD behaviour. We design a small experiment to verify that the choices of source contents greatly affect the final experiment result:

Instead of 20 source content, 10 out of 20 source contents' RD statistics are picked randomly and the encoder BD-rate performance for them are analyzed. Three encoders, HEVC, VP9 and AVS2, are selected for the encoder performance comparison. Below is a table summarizing their performance rank. The experiment is repeated 100 times with random selections of the 10 source contents.

Table 3.4: Percentage of Encoders' Rate Saving Ranking Order for Different 10 Source Contents (Experiment Repeated 100 Times)

Ranking	Percentage
HEVC>VP9>AVS2	30%
VP9>HEVC>AVS2	26%
HEVC>AVS2>VP9	14%
VP9>AVS2>HEVC	16%
AVS2>VP9>HEVC	8%
AVS2>HEVC>VP9	6%

The results in Table.3.4 indicate that the choice of source contents plays an important role in the final result of compression applications. If the number of source contents is decreased to about 10, as is the case of most video codings standard development in the past 3 decades. the encoder performance would vary dramatically depending on the selected contents, which can be inevitably and unconsciously biased due to researchers' preference. Even though the problem of inconsistent result due to source contents can be solved by increasing the number of source videos, the number has to be limited in practice. As discussed in previous sections, quality database's size is limited due to the high cost such as number of participants in subjective experiments and the availability of computing resource. The process of quality database construction begins with collecting a large number of candidates for the source contents, as shown in Fig.3.4. Researchers have to examine not only the quality of the collected videos one after another but also

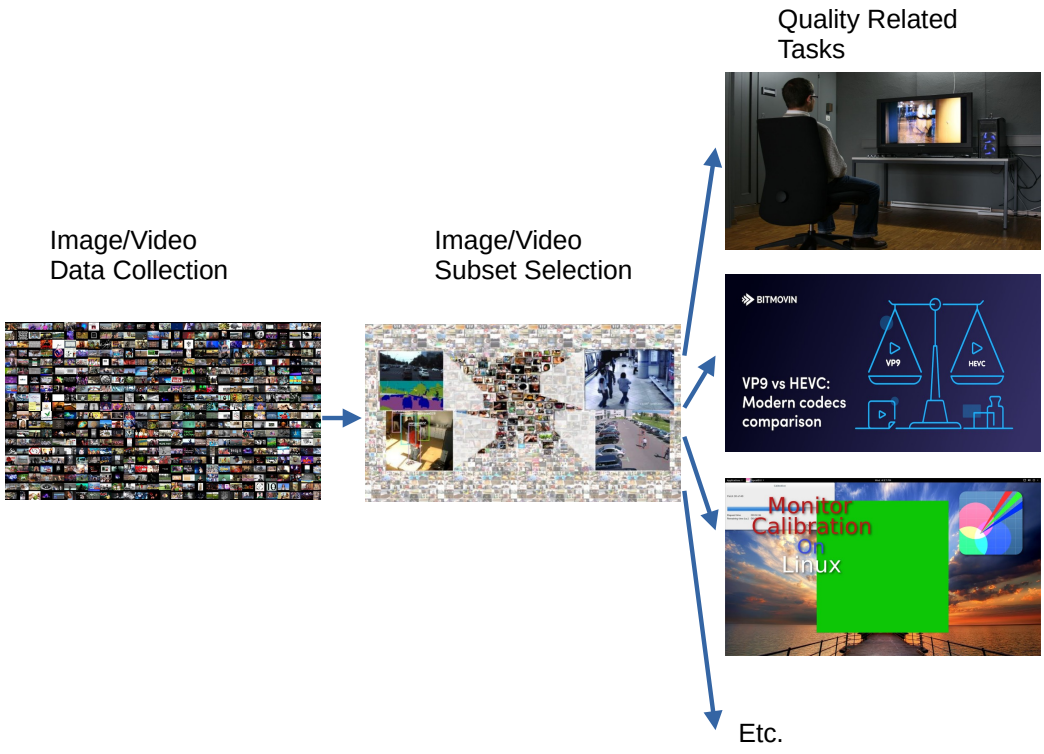


Figure 3.4: Quality database construction process

considering many factors such as content types, Colourfulness (CF), and SI etc.. With all collected characteristics taken into consideration, the task of source content selection considering information from thousands of visual contents is subjective and extremely difficult. Therefore, it is necessary to design an automated source content selection method based on the visual content characterization model so that representative source contents can be objectively selected for visual quality dataset.

3.5 Conclusion

We conduct an independent and so far the most comprehensive subjective evaluation and performance analysis, specifically on popular and emerging video encoders (AVC, HEVC, VP9, AVS2, and AV1) with video content of diverse resolutions and bitrates. The five video encoders are evaluated across 20 source 4K contents from the view points of content dependency and resolution adaptation. The testing results have been made publicly available to facilitate future video coding and VQA research. Furthermore, we comparatively analyze the encoder performance through subjective experiment and find out that the increased flexibility enables recent video encoders to have better efficiency in RD sense but also brings the problem of increased coding complexity.

Based on the observations from database construction and encoder performance analysis, current quality control mechanism implemented in most video encoders hinders the distortion level selection, which is often carried out based on inefficient visual check procedures. A direct modelling between encoder parameters such as Quantization Parameter (QP) against HVS inspired metric such as SSIM would speed up the database construction process. What is more, the better quality control mechanism would alleviate the problem of quality misalignment across different RD curves. Therefore, in Chapter 4, we will introduce the RD analysis motivated encoding parameter selection, which will demonstrate the effectiveness of using encoding RD analysis for content characterization for compression applications.

We also observed that the widely used visual features, the SI and TI, are not reliable in characterizing source contents for compression quality database. Shown by the example of the two contents sharing the same or similar SI and TI scores, their drastically different RD behaviours justify the motivation and main theme of the thesis, which aims to utilize encoding RD for visual content characterization for the compression applications.

Moreover, the small scale encoder performance comparison experiment using randomly selected content demonstrates the importance of source content selection. The number of source contents included in a image/video compression database is largely limited by the high cost of computing resources and subjective experiments. To overcome these limitations, it is critical to develop a systematic approach so as to select the most representative

videos or images that exhibit diverse and representative characteristics, best reflecting content variations and human visual quality experience. Therefore, in Chapter 5, we will propose an objective source visual content selection method, which is based on encoding RD analysis and submodular optimization.

Chapter 4

Quality Control for Visual Coding by Eigen Analysis of Generalized Quality Parameter Functions

While pushing the rate-quality (RD) limit to the next level will always be a paramount goal of video/image compression models, precise control in terms of both rate and quality is the necessary cornerstone for real world applications. Recent years have witnessed great advancement of accurate rate control method for video encoders, whereas quality control that based on the accurate video quality and encoding parameter modelling is less visited. In this work, the generalized function space of quality-encoding parameter inspired by Rate-Distortion (RD) analysis is constructed and analyzed. Then an eigen analysis approach is proposed for the modelling of image/video quality against the encoding parameter, which is named as generalized quality parameter (GQP) model. The theoretical function space is defined and proved to be a convex set in a Hilbert space, which inspires a computational model of GQP function and a method of sparse measurements parameter estimation. Two large-scale databases, one for videos and the other for images, are used to demonstrate the idea through experiments. With the computational model and the sparse measurements method, the GQP function of a specific video/image can be reconstructed accurately from only a few queries, which significantly outperforms the current widely used

empirical estimation methods both in accuracy and efficiency.

4.1 Motivations and Related Work

RD trade-off plays a key role in visual compression models. For modern video encoding method such as HEVC [85] and VVC [86], with more flexible coding parameters, such as increased number of intra-prediction modes and quad-tree coding unit structure, the introduction of rate-distortion optimized encoding parameter selection has enabled HEVC and VVC to out-perform H264 [36] by a large margin in terms of RD performance [43, 87]. On the other hand, in the area of image compression, the rate-distortion trade-off is incorporated as the optimization goal in many recent End-to-End (E2E) optimized learning based methods [88, 89, 53, 90, 91]. While pushing the RD limit to the next level will always be a paramount goal of modern video/image encoders, precise controlling in terms of both rate and quality is necessary in real world applications of video/image compression such as video transmission under limited bandwidth and targeting constant quality of experience for video viewers. With the improvement in RD performance achieved in both video and image compression area, RD optimization casts a great challenge on rate control and quality control. The rate/quality against encoding parameters relationship across different source video or image contents is becoming increasingly sophisticated because of the more flexible yet much more complicated encoding parameter combinations in video compression methods and the Lagrangian multiplier introduced for RD trade-off control in learning based image compression methods [92, 93].

Rate control, aiming for achieving target data rate with a small variation across different parts of the visual content and meanwhile maintaining the best possible quality, has always been a frequently visited problem in video/image compression area [94, 95]. Controlling parameters, such as Quantization Parameter (QP), modelled against rate, combined with bit allocation methods plays a key role in rate control of modern video encoders such as VVC, HEVC and H264 and has gained popularity due to its ability in controlling the final encoded video's bitrate in a simple way. However, quality control, aiming for finding the most suitable encoding parameters for a specific quality level, has not been paid much

attention in the research field. It seems an easy solution to replicate the rate control models by constructing an exponential or hyperbolic function between quality and encoding parameters [43]. However, it is impossible to re-allocate the quality across frames as rate allocation method does. It is also problematic to take the average quality across all frames and treat it as the final quality score for the video due to the complication of human observer's sensitivity to quality variation[96]. Moreover, it is common to see modelling between QP and data-rate fail for some specific contents due to extremely low/high bitrate requirement, where the bit allocation algorithms compensate the bits by discarding frames or padding meaningless bits in the encoded bitstream. The same philosophy of reallocating or compensating bits cannot be applied if we expect the quality of the encoded video to be constant on a specific quality level. Constant rate factor (CRF) encoding mode implemented in x264/x265 video encoders is one of the few efforts targeting a constant quality level for the encoded videos. [51, 36] However, according to the studies of CRF and video encoder comparison [47, 44, 97], even though the quality is seemingly closer to a constant level throughout the whole video sequence, the actual quality level in terms of human vision perception driven metrics such as SSIM [98], SSIMplus [8] or VMAF [99] can vary drastically for different content at the same CRF value. The video encoders' CRF method often fail the task of constant quality when measured by human perceptual visual quality.

On the other hand, in recent successful learning based E2E image compression models [88, 89, 53, 90, 91], researchers focus on achieving better rate-distortion performance while neglecting the rate/quality control targeting tasks for real world applications. The established vision encoding methods on rate/quality control, which are mostly based on heuristic models such as exponential or hyperbolic functions and hand-tuned parameters[43], would fail due to the totally different learning based frameworks where a single hyper-parameter λ in cost function may be used to control the trade-off between rate and quality. Therefore, without an accurate mapping between human perceived quality and encoding parameters, such as QP in video coding and λ in E2E image coding, the final human perceived quality of encoded video/image for a specific content is uncontrollable.

Inspired by previous work in modeling generalized rate distortion (GRD) functions [1], we propose an eigen analysis approach of generalized quality parameter (GQP) model in

tackling the problem of quality and encoding parameter relationship modelling. Firstly the theoretical functional space $\mathcal{W}$ of the GQP function is defined by analyzing its mathematical properties, which is lacking in the current video/image encoding research area. It can be shown that with the monotonic relationship between quality and controlling parameter, $\mathcal{W}$ is a convex set in a Hilbert space, inspiring a computational model of the GQP function, and a method of estimating model parameters from sparse measurements. To demonstrate the feasibility of the idea, a large-scale database of real-world video GQP functions is collected, which turns out to live in a low-dimensional subspace of $\mathcal{W}$. Combining the GQP reconstruction framework and the learned low-dimensional space, a low-parameter eigen GQP method is create to accurately estimate the GQP function of a source video content from only a few samples. Experimental results on the database show that the learned GQP method significantly outperforms widely used empirical estimation methods both in terms of accuracy and efficiency. Lastly, we demonstrate the promise of the proposed model in E2E image quality control parameter modeling and x265 video Human Visual System (HVS) quality control.

The chapter is organized in the following order. In section 4.2, the theoretical space of the encoding parameter and quality will be established by showing the mathematical properties of the GQP functions and concluding with the convexity of the space $\mathcal{W}$. In section 4.3, the approximation framework for GQP functions in functional space $\mathcal{W}$ is discussed in detail. In section 4.4, the database collection and the eigen GQP (eGQP) model for the collected GQP functions will be discussed. In section 4.5, the effectiveness of the proposed eGQP model will be demonstrated in quality control using x265 video encoder [51]. In section 4.6, the application of eGQP with E2E image compression [88] quality control will be demonstrated. In section 4.7, conclusions will be drawn.

4.2 Theoretical Space of GQP Functions

Defining the theoretical space of GQP functions help us better understand the model of these functions. Guided by the defined theoretical space, the form of the model are obtained, together with the constraints these functions must satisfy. The analysis begins

by stating the assumptions of the desired GQP functions.

The first assumption is that the domain of the GQP functions is a compact set Ω. The closed and bounded interval in parameter space is a typical setting of Ω, *i.e.*, $x \in \Omega = [x_{\min}, x_{max}]$, where x is the encoding parameter that can represent the quantization step in video encoders or hyper-parameter λ in E2E image encoders. Based on the type of practical applications, x_{min} and x_{max} can be easily determined. For example, in the following video encoder experiment, the maximum QP value 51 and minimum QP value 0 are taken as the maximum and minimum. When QP equals to 0, the video is losslessly encoded, which means $f(0) = z_{max}$, where z_{max} denotes the best quality in terms of the quality metric. When QP equals to 51, the encoded video is most degraded, which means $f(51) = z_{min}$, where z_{min} denotes the worst quality the encoder can achieve for the specific content in terms of the quality metric. Without loss of generality, we normalize the range of GQP functions such that $z_{min} = 0$ and $z_{max} = 100$ [75, 100]. It is worth noting that for video encoding application, in order to make the quality score normalized within the the range of $[z_{min}, z_{max}]$, where $z_{min} = 0$, one extra pseudo parameter is added that will make $z_{min} = 0$, which means QP = 52 is set as the x_{max} for $z_{min} = 0$.

The second assumption is that the GQP functions are continuous, *i.e.*, $f \in C(\Omega)$. Even though the quantization steps are discretized in real world applications such as HEVC and H264, it can always be observed that successive change in quantization steps of video encoders leads to the gradual transitions in perceptual quality in many subjective user studies [101, 102, 103]. Moreover, the gradual changes of Lagrangian multiplier in E2E image encoders results in the changes of rate-quality trade-off. Thus the continuous assumption holds for both scenarios. Some sample GQP functions for video and image applications are shown in Fig.4.1 and Fig.4.2, where the continuous trend can be observed.

The third assumption is that the GQP functions are monotonic. In video encoders, some of the key tuning parameters such as QP or CRF are designed to strictly follow the monotonic relationship against the perceived quality of the encoded video [47, 44]. For example in HEVC video encoder, with QP getting larger, the quantization step increases until the QP value reaches maximum, corresponding to the largest quantization step, leads to the worst quality of encoded video. Moreover, according to RD theory, the second order relationship of quality vs. rate brings a monotonic relationship of controlling parameter λ

against quality. Therefore, the monotonic GQP functions assumption applies to E2E image compression models as well. The sample QP-Quality curves as well as λ-Quality curves shown in Fig.4.1 and Fig.4.2 demonstrate the monotonic relationship between quality and encoding parameters. Without loss of generality, we assume f increases monotonically with the increase of encoder control parameter. If f is monotonically decreasing, such as the case of QP and CRF in video encoders, the function can be simply replaced with $z_{max} - f$, where z_{max} is the maximum value of the quality metric.

Under the aforementioned assumptions, the space of GQP functions is defined as:

$$\begin{aligned}\mathcal{W} \coloneqq \{f : \mathbb{R} \mapsto \mathbb{R} | f \in C(\Omega); f(x_{\max}) = 100; \\ f(x_{\min}) = 0 \text{ and } f(x_a) \leq f(x_b), \forall x_a \leq x_b\}.\end{aligned} \tag{4.1}$$

The two equality constraints in $\mathcal{W}$ form an affine space $\mathcal{H}_1$, which can be described as a linear subspace

$$\begin{aligned}\mathcal{H}_0 \coloneqq \{f : \mathbb{R} \mapsto \mathbb{R} | f \in C(\Omega); \\ f(x_{\min}) = 0\}\end{aligned} \tag{4.2}$$

translated by any function $f_0 \in \mathcal{H}_1$. Formally, we may express the relationship between $\mathcal{H}_1$ and $\mathcal{H}_0$ by

$$\mathcal{H}_1 = f_0 + \mathcal{H}_0, \forall f_0 \in \mathcal{H}_1. \tag{4.3}$$

The inequality constraints form a convex cone

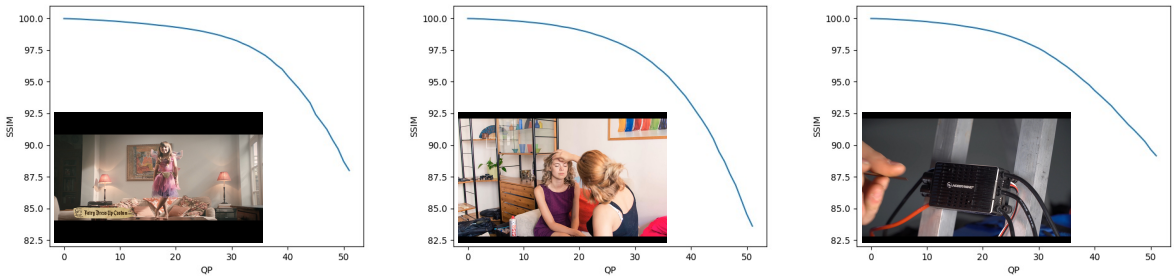
$$\mathcal{V} \coloneqq \{f : \mathbb{R} \mapsto \mathbb{R} | f(x_a) \leq f(x_b), \forall x_a < x_b\},$$

where it can be shown that $\forall \alpha, \beta \geq 0$ and $v_0, v_1 \in \mathcal{V}$, $\alpha v_0 + \beta v_1 \in \mathcal{V}$.

Finally, it can be concluded that the theoretical space $\mathcal{W}$ can be described as the intersection of the affine space $\mathcal{H}_1$ and the convex cone $\mathcal{V}$:

$$\mathcal{W} = \mathcal{H}_1 \cap \mathcal{V}. \tag{4.4}$$

Thanks to the convexity of $\mathcal{H}_1$ and $\mathcal{V}$, it can be readily shown that $\mathcal{W}$ is a convex set.



(a) (b) (c)

Figure 4.1: Samples of GQP curves for different video content compressed by x265 video encoder.

4.3 GQP Function Approximation Framework

With the relations $\mathcal{H}_1 = f_0 + \mathcal{H}_0$ and $\mathcal{W} = \mathcal{H}_1 \cap \mathcal{V}$ at hand, the infinite-dimensional space $\mathcal{W}$ can be parameterized. Firstly, it can be concluded that $\forall h \in \mathcal{H}_0$, and h is square-integrable because h is a continuous function defined over a compact set shown by Eq. (4.2). Therefore, the inner product of space $\mathcal{H}_0$ is

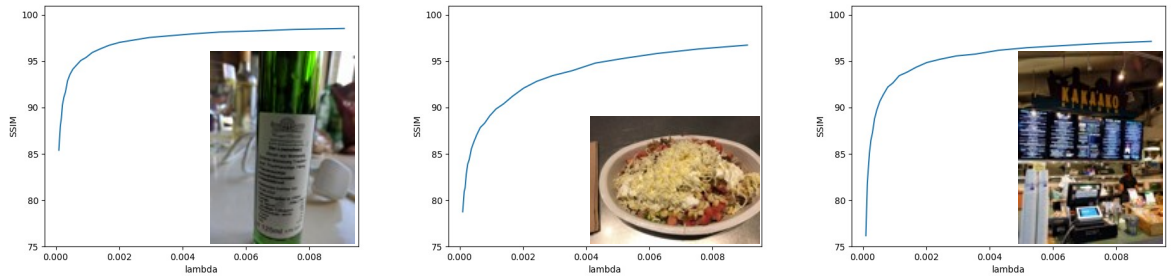
$$\langle h, g \rangle \coloneqq \int_{\Omega} h(x)g(x)dx, \forall h, g \in \mathcal{H}_0, \tag{4.5}$$

and an induced metric can be defined as

$$d_2(h, g) \coloneqq \left[\int_{\Omega} |h(x) - g(x)|^2 dx \right]^{\frac{1}{2}}, \forall h, g \in \mathcal{H}_0.$$

With metric d_2 defined and including the limits of all Cauchy sequences that belong to the functional space, $\mathcal{H}_0$ is completed. Since the space of all square-integrable functions defined on Ω, denoted as $\mathcal{L}_2(\Omega)$, is a Hilbert space with Eq. (4.5) being the inner product operation, $\mathcal{H}_0$ is a dense subset of $\mathcal{L}_2(\Omega)$ [104].

Then we can know $\mathcal{H}_0$ is separable, as polynomial functions form a dense countable subset of $\mathcal{H}_0$ [104]. It can be concluded that an orthonormal basis $\{h_n, n = 1, 2, 3, \cdots\} \subset$



(a) (b) (c)

Figure 4.2: Samples of GQP curves for different image content compressed by E2E models.

$\mathcal{H}_0$ exists and spans $\mathcal{L}_2(\Omega)$:

$$h = \sum_{n=1}^{\infty} c_n h_n, \quad \forall h \in \mathcal{L}_2(\Omega) \tag{4.6}$$

As a result, any GQP function $f \in \mathcal{W}$ can be expressed as a linear combination of $\{h_n\}$:

$$\exists \{c_n\}, \text{such that } f = f_0 + \sum_{n=1}^{\infty} c_n h_n, \forall f \in \mathcal{W}. \tag{4.7}$$

The parameterization in Eq. (4.7) is equivalent of a series of approximation models such as the N-th order approximation model:

$$\tilde{f} = f_0 + \sum_{n=1}^{N} c_n h_n, \tag{4.8}$$

Better approximations may be achieved using larger values of N. A systematic way of estimating a GQP function from samples is described as below:

Following the N-th order model in Eq. (4.8), an N-dimensional approximation of $\mathcal{W}$ can be defined as:

$$\tilde{\mathcal{W}}_N \coloneqq \left\{ f \middle| f = f_0 + \sum_{n=1}^{N} c_n h_n, f \in \mathcal{V} \right\}. \tag{4.9}$$

The approximation space $\tilde{\mathcal{W}}_N$ is a subset of $\mathcal{W}$ as $\{h_n\} \subset \mathcal{H}_0$, meaning that any element in $\tilde{\mathcal{W}}_N$ is a valid GQP function. Therefore, estimating a GQP function is equivalent to

finding the optimal element in $\tilde{\mathcal{W}}_N$ that best fits given samples. Because of $\tilde{\mathcal{W}}_N$ is a closed convex set, we can formulate GQP function estimation as a projections-onto-convex-sets (POCS) problem. Given a set of quality-parameter pairs $\{f(x_i) = z_i, i \in \mathcal{I}\}$, where $\mathcal{I}$ denotes the index set, we aim to solve

$$\begin{aligned} \arg\min_{\{c_n\}} \quad & \sum_{i\in\mathcal{I}} |z_i - f_0(x_i) - \sum_{n=1}^{N} c_n h_n(x_i)|^2 \\ \text{s.t.} \quad & f_0 + \sum_{n=1}^{N} c_n h_n \in \mathcal{V}. \end{aligned} \tag{4.10}$$

Then optimal coefficients $\{c_n^*\}$ can be plugged into Eq. (4.8) to obtain the estimated GQP function.

4.4 eGQP Model

Thanks to the GQP function approximation framework, arbitrary orthonormal basis can be chosen to approximate the GQP functions. For example, h_n can be chosen to be a second or third degree polynomial basis or trigonometric basis such as half-sine which is orthonormal in nature. However, the predefined basis are not adaptive that they may not capture the large variation in data manifold and therefore result in a large number of basis functions to achieve acceptable accuracy. In order to find an efficient set of basis, the eigen basis is proposed which is constructed based on principal component analysis. In this section, the data driven eigen basis is introduced following the description of the large scale GQP function database constructed through x265 video encoder. The method makes it possible to reconstruct GQP function through only a few data adaptive eigen basis with a satisfactory approximation accuracy using only a few sparse samples.

4.4.1 Optimal Basis of Real-World GQP Functions

GQP Function Database

A large scale GQP function database is necessary in two ways:

- The data-adaptive eigen basis can be obtained through a principal component analysis process of the GQP function database that contains a set of parameter-quality curves, for example, QP-SSIM curves in our video quality control experiment.
- The database can also be used as a test bed for comparing the performance of different basis used under the approximation framework described in Section III.

Thanks to the great effort paid in constructing the video database described in [1], the 1000 1920x1080 resolution pristine source videos are utilized as video database to construct the GQP function database. As shown in Fig 4.3, the video database covers a broad range of contents. The processing of source videos is as follows. Firstly, due to the fact that content varies little during the playback of each source video, first one second of each source content is extract from the original ten-second video. Then the 1000 one-second videos are encoded with the main profile of x265 using the constant QP mode that range from $QP = 0$ to $QP = 51$. In the end, 52000 encoded one-second video clips are obtained in total. Secondly, the frame-level SSIM score is computed for each of the encoded video clips and a GQP function curve for the same content frame is obtained. The total number of GQP curves for the video database is over 1.5 million.

It should be noted that other perceptual quality metrics can also be used, the reason SSIM is selected is due to its wide accessibility as many modern video encoders have the built-in SSIM module. As mentioned in previous sections, due to the complexity of the source contents, the lowest achievable quality varies across different contents when the QP value is maximum at 51. Therefore, one extra value pair of QP-quality of $(52, 0)$ is padded so that the the lowest quality and highest quality, which are achieved at $QP = 52$ and $QP = 0$, respectively, are aligned for each source content and the condition in Eq. (4.1) is met.

Eigen Basis for Real-World GQP Functions

Since the goal is to find the optimal set of basis that approximates the real-world GQP functions, the approximated function can be firstly defined for a specific real-world GQP

Figure 4.3: Sample frames of source videos in the Waterloo GRD database. All images are cropped for neat presentation.[1]

function $\mathbf{f}_m$ using Eq. (4.8) as:

$$\tilde{\mathbf{f}}_m \coloneqq \mathbf{f}_0 + \sum_{n=1}^{N} \langle \mathbf{f}_m - \mathbf{f}_0, \mathbf{h}_n \rangle \mathbf{h}_n,$$

where m denotes the m-th function and it is approximated by N of the optimal basis. Therefore, the approximation error is given by

$$\mathcal{E}[\mathbf{f}_m] \coloneqq \left| \mathbf{f}_m - \left(\mathbf{f}_0 + \sum_{n=1}^{N} \langle \mathbf{f}_m - \mathbf{f}_0, \mathbf{h}_n \rangle \mathbf{h}_n \right) \right|_2, \tag{4.11}$$

which is a Euclidean norm of a vector. It should be noted that a discrete version of the basis function is used as $\mathbf{h}_n$ of the h_n defined in Eq. (4.8). With M empirical GQP functions in the video GQP database, the optimal orthonormal basis is therefore obtained by minimizing the average approximation error:

$$\begin{aligned} \underset{\mathbf{f}_0, \{\mathbf{h}_n\}}{\arg\min} \quad & \frac{1}{M} \sum_{m=1}^{M} \left| \mathbf{f}_m - \mathbf{f}_0 - \sum_{n=1}^{N} \langle \mathbf{f}_m - \mathbf{f}_0, \mathbf{h}_n \rangle \mathbf{h}_n \right|_2^2, \\ \text{s.t.} \quad & |\mathbf{h}_n|_2^2 = 1, \quad n = 1, \cdots, N, \\ & \langle \mathbf{h}_n, \mathbf{h}_{n'} \rangle = 0, \quad n, n' \in \{1, \cdots, N\}, n \neq n'. \end{aligned} \tag{4.12}$$

It can be easily shown that the optimal $\mathbf{f}_0^*$ equals to the mean of M GQP functions when $N = 0$, which can be proved by the convexity of $\mathcal{W}$. It can also be proved that the problem (4.12) is essentially the principal component analysis (PCA) as N goes above 0. According to the definition of PCA, the n-th optimal component $\mathbf{h}_n^*$ is the eigenvector associated with the n-th largest eigenvalue of the empirical covariance matrix of $\mathbf{f}_m$. The space $\mathcal{W}$'s optimal N-dimensional approximation can be achieved by the span of the first N eigenvectors and $\mathbf{f}_0$ as well.

In order to demonstrate the effectiveness of the principal components in explaining the space $\mathcal{W}$, the plot of the cumulative energy with respect to number of components is shown in Fig 4.4 on the video GQP database. With only 3 components, over 99.5% of the energy is covered, which suggests that most real-world GQP functions lie in a much lower dimensional space and that the eGQP models with only a few parameters would work well. The $\mathbf{f}_0^*$ plus the first 7 components are plotted in Fig 4.5 to help gain a better understanding

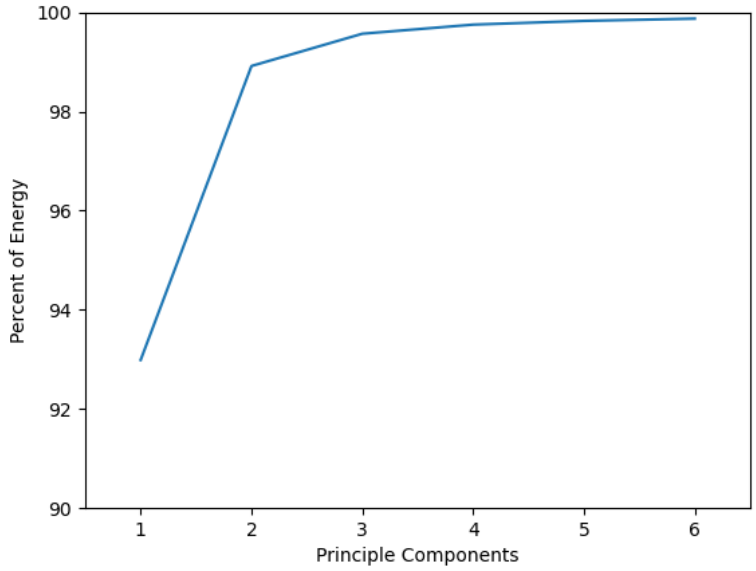


Figure 4.4: The percentage of the energy explained by the span of the first 6 principal components.

about the shapes of the eigen GQP functions. Two observations can be concluded from the figure. Firstly, the $\mathbf{f}_0^*$ is the smoothest, while the second to the fourth component are increasingly oscillatory. This implies that the perceptual quality of a video representation is positively correlated with its neighboring representations in general. Second, all the principal components exhibit the greatest magnitudes in regions with high QP, indicating the perceptual quality varies drastically across different contents with the same parameter settings in low quality region.

4.4.2 eGQP Model Estimation from Sparse Samples

Since over 99.5% of the energy can be efficiently represented by the subspace of the first 3 components, the parameters of the eGQP model can be accurately estimated through inserting the learned mean component $\mathbf{f}_0^*$ and the principal components $\{\mathbf{h}_n^*\}$ into the POCS problem. Constraints can be approximated as a set of linear inequalities, which

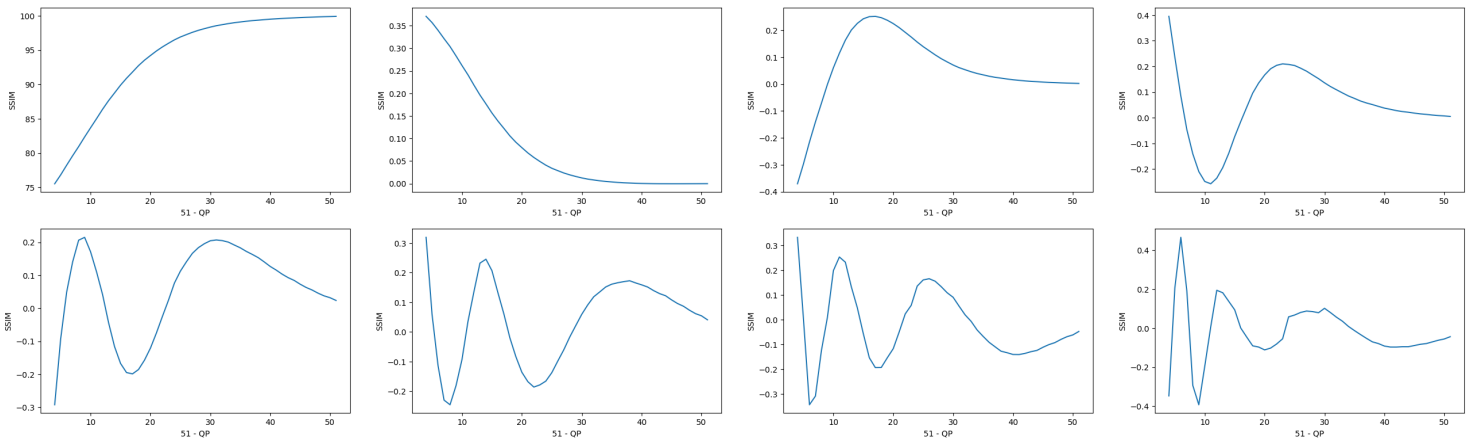


Figure 4.5: The mean and first seven principal components of the real-world x265 GQP functions.

makes the Problem (4.10) practically solvable. The Eq. (4.8) can be rewritten in matrix form as:

$$\tilde{\mathbf{f}} = \mathbf{f}_0 + H_N^* \mathbf{c}, \tag{4.13}$$

in which $H_N^* \coloneqq [\mathbf{h}_1^*, \mathbf{h}_2^*, \cdots, \mathbf{h}_N^*]$ and $\mathbf{c} \coloneqq [c_1, c_2, \cdots, c_N]^T$. The inequality constraint can be translated as the discrete form of

$$D_x \ \tilde{\mathbf{f}} \geq 0. \tag{4.14}$$

where D_x is the first order difference along the x-axis. Then (4.13) can be substituted into (4.14) and the following can be obtained

$$- \ D_x \ \mathbf{H}_N^* \mathbf{c} \leq \ D_x \ \mathbf{f}_0, \tag{4.15}$$

which imposes a linear constraint on the coefficients $\mathbf{c}$. Therefore, finding optimal $\mathbf{c}^*$ is equivalent to a quadratic programming problem, which can be solved by convex optimization tools, such as OSQP [105]. In the end, by substituting $\mathbf{c}^*$ into Eq. (4.8), the best eGQP model is obtained that fits known samples with the least squared errors.

4.5 Experiments

The first round of experiments are based on the constructed video GQP database. In this section, firstly the approximation capability of the GQP framework is demonstrated by quantitatively comparing the accuracy of restoring GQP functions using different numbers of basis functions. The performance of three different basis functions namely, polynomial, trigonometric and eigen basis will also be compared. The second experiment demonstrates the proposed framework's capability of GQP function reconstruction from sparse samples. Since only a few number of eigen basis recovers over 99.5% of the GQP space energy, a GQP function can be easily recovered with only a few samples of the actual GQP function, therefore making it possible for real world application of constant perceptual quality control. Following the sparse sampling experiment, a quantitative analysis is done on the question of how many samples/basis will be good enough for a acceptable recovery of GQP function. Then effectiveness of the monotonicity constraints assumption on preventing the over-fitting problem will be shown. The effectiveness of proposed eGQP framework on another video quality metric is also demonstrated using VMAF as the quality metric with x265 encoded videos. Lastly a real world application using x265 video encoder for constant visual quality control is conducted.

4.5.1 Approximation Capability Comparison

In Table 4.1, we show the approximation capability of the eGQP functions when the number of basis increases by using the root mean square error (RMSE) and l^∞ norm as the two metrics. The RMSE is used for measuring the average error across the data points for GQP curves. The l^∞ norm is used for measuring the extreme irregularity of the reconstructed function against the original GQP curves. Moreover, the worst case scenario for the two metrics are included in the table. The eGQP basis are obtained using the PCA method on training split of the constructed video database. The training and testing ratio is 8:2. The proposed method's approximation capability is demonstrated in Table 4.1. Since the purpose is to show the reconstruction capability, the results obtained from the training database are used. It should be noted that the SSIM quality metric used in the experiment

is scaled to $[0, 100]$ range from $[0, 1]$ for comparison convenience.

Table 4.1: Mean and worst performance of eGQP on the training set with different numbers of basis functions

N	RMSE		l^∞ error	
	Mean	Worst	Mean	Worst
0	4.41	21.27	11.40	53.54
1	0.66	7.03	1.75	13.27
2	0.21	1.50	0.54	5.15
3	0.10	0.73	0.24	2.12
4	0.06	0.38	0.18	1.24

As discussed in Section II, the GQP approximation framework can choose any linear basis such as polynomial or trigonometric functions. Therefore, in Table 4.2 we compare the two aforementioned bases against the proposed eigen basis using an increasing number of basis until 4, for which the eigen basis covers over 99.9% of the energy for GQP space basis. As shown in the table, when compared with the other two basis, the eigen basis gain the largest accuracy improvement by increasing the number of basis in terms of l^∞, which compares the largest deviation for a specific GQP function. It can also be observed that as the number of basis increases, the other two basis's accuracy only increases slowly. The phenomena can be explained by the data-driven nature of the eigen basis, which efficiently captures information needed to explain the GQP space. It is worth noting that the worst case performance for polynomial and trigonometric basis doesn't change at all, which is drastically different from its counterpart of eigen basis. This implies that the predefined basis functions deviate from real world GQP functions by a large margin and the data-driven eigen basis well captures the characteristics of real world GQP functions.

4.5.2 GQP Function Reconstruction from Sparse Samples

An accurate modelling of the encoding parameter against quality would guide the selection of encoding parameters so that the target constant quality can be achieved. With a given

Table 4.2: l^{∞} error of GQP models with different basis functions on the test set

N	Polynomial		Trigonometric		Eigen	
	Mean	Worst	Mean	Worst	Mean	Worst
0	*11.047*	**36.783**	*11.047*	**36.783**	*11.047*	**36.783**
1	11.003	36.783	11.007	36.783	*1.696*	**9.861**
2	10.964	36.783	10.961	36.783	*0.555*	**2.454**
3	10.892	36.783	10.892	36.783	*0.235*	**0.936**
4	10.889	36.783	10.891	36.783	*0.173*	**0.722**

model, the more samples obtained, the more accurate of the reconstruction results will be. For most encoding parameter relationship models currently used, such as exponential or hyperbolic models[43], the accuracy improvement comes at the expense of obtaining parameter-quality sampling points through multi-pass encoding, which often brings unaffordable complexity burden to the encoding pipelines that are time-critical. Therefore it is desirable for the encoding parameter to be selected in a time efficient manner. Most importantly, the number of samples needed to achieve the best compromise between accuracy and cost has never been analyzed. In this section, we conduct an analysis experiment on GQP function reconstruction from sparse samples. In addition to the proposed eGQP framework, the models we used for comparison include hyperbolic model and exponential model, which are common in x265 and x264 video encoders. In order to maximize the accuracy of the reconstructed functions, the sampling strategy plays an essential role. In this work, we select an information-theoretic sampling method[9], which minimizes the uncertainty of the function by generating a fixed sample sequence.

A total of 3 GQP functions are tested including the two widely used models, exponential and hyperbolic[43], and the GQP frameworks with eigen basis. The results shown in the Table 4.3 are tested using the testing set of GQP database. The training set, which takes 80% of the whole dataset, is used for training the eigen basis. In order to conduct a fair comparison, the number of basis for eGQP framework used is 2, which is equal to the number of free parameters of exponential and hyperbolic functions. In Table 4.3, it is obvious that the eGQP method outperforms the exponential and hyperbolic function by a

Table 4.3: RMSE of GQP models with different sample numbers

S	Hyperbolic [43]		Exponential [43]		eGQP	
	Mean	Worst	Mean	Worst	Average	Worst
2	6.63	100.80	4.19	23.52	*0.28*	**2.11**
4	3.70	29.35	2.47	11.40	*0.23*	**1.74**
8	2.99	20.71	2.13	9.29	*0.22*	**1.66**
20	2.27	12.99	1.82	7.85	*0.21*	**1.68**
40	1.88	9.93	1.61	7.05	*0.20*	**1.56**

large margin. This is because the competing methods presume a fixed function form that fail to explain the GQP space accurately. Moreover, in the case of only two samples, the eGQP method can recover the original function with the lowest error, which builds the foundation towards accurate encoding parameter selection for constant perceptual quality control. The error of the eGQP method is so low that the subjective quality difference can be ignored [75] according to previous studies [106, 107], which are often regarded as indistinguishable to human observer in subjective studies. Because the number of samples required is limited, which is only 4 in the previous experiments, the proposed eGQP framework offers a great option to achieve the balance between time efficiency and prediction accuracy. In real world applications, the sampling process can be implemented concurrently and the time efficiency is achieved.

4.5.3 Influence of Varying Number of Bases on eGQP

For experiments conducted in previous section, the number of basis functions is fixed to 2 and the number of samples used vary between 2 and 40. In order to evaluate the contribution of varying number of bases to the accuracy, we conduct the experiment of varying number of both bases and sparse samples. The results are shown in Table VI. It can be seen that by increasing the number of bases with the number of samples, the performance of eGQP can further improve, which is supported by the fact that each basis function explains the GQP space further more. The results shown in the following sections

are all obtained using variable numbers of samples and basis.

Table 4.4: Mean and worst performance of eGQP when the number of basis vectors is equal to the number of samples

N/S	RMSE		l^∞ error	
	Mean	Worst	Mean	Worst
1	0.70	6.90	1.98	13.69
3	0.19	0.90	0.40	1.98
5	0.06	0.36	0.24	2.06
8	0.03	0.18	0.18	1.24
30	0.01	0.07	0.17	0.71

4.5.4 Importance of Monotonicity Constraint

One of the most important assumptions of the GQP framework is the monotonicity of encoding parameter against the quality of image/video. In order to demonstrate the importance of the constraint, we show the accuracy of eGQP without monotonicity constraint in Table VII. It can be seen that even though it has similar accuracy when the number of samples is low, the average error increases when the number of samples increases, which means the eGQP without monotonicity constraint is prone to over-fitting when the number of samples/basis increases. The monotonicity assumption works as a prior knowledge in the proposed framework so that over-fitting can be avoided.

4.5.5 eGQP with Different Quality Metrics

The quality metric SSIM is used for experiments conducted so far due to the ease of access for most video encoders, which comes as built-in feature in most popular and publicly available video encoders. Since VMAF [99] is a quality metric that are designed specifically for VQA tasks and is open source, we use VMAF in this experiment to demonstrate the generalization capability of proposed eGQP framework on different VQA metrics. The

Table 4.5: Mean and worst performance of eGQP without monotonicity constraints

N/S	RMSE		l^∞ error	
	Mean	Worst	Mean	Worst
1	0.70	6.90	1.98	13.69
3	0.19	0.90	0.40	1.98
5	0.06	0.36	0.24	2.06
8	0.04	0.30	0.18	1.24
30	0.04	0.18	0.23	1.05

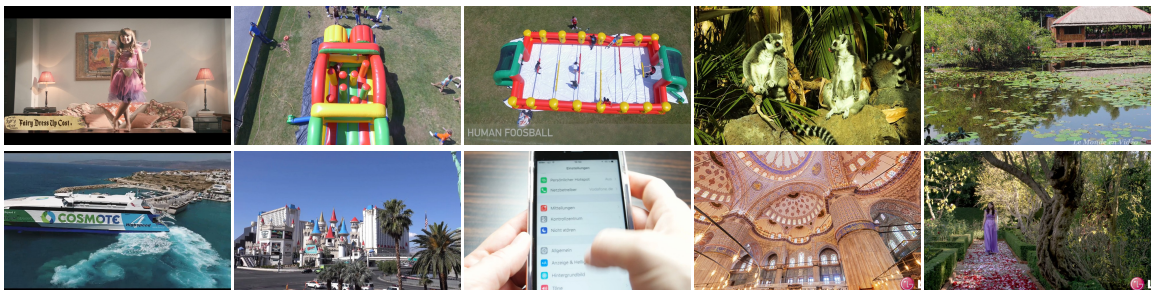


Figure 4.6: Sample frames of source videos for eGQP target constant quality experiment

results are obtained from the same testing set with the uncertainty sampling strategy introduced in previous section. Three observations should be noted in Table 4.6. First it can be seen that eGQP enjoys the same trend of accuracy gain similar to the SSIM results in Table 4.4 as the number of samples/basis increases. Secondly for the same number of samples/basis, VMAF's RMSE and l_∞ error are larger than that of SSIM, which is due to the fact the range of $[90, 100]$ is the effective range of SSIM for most of the contents while the VMAF occupies the range of $[0, 100]$ for video contents. Thirdly, the low RMSE and l_∞ error enables the video encoders to achieve the target constant quality in terms of VMAF video quality metric.

Table 4.6: Mean and worst performance of eGQP on GQP functions measured by VMAF

N/S	RMSE		l^∞ error	
	Mean	Worst	Mean	Worst
1	1.24	17.33	3.24	22.12
3	0.59	2.13	1.54	5.34
5	0.36	1.57	1.48	7.08
8	0.23	1.13	1.15	6.89
30	0.08	0.40	0.86	6.43

4.5.6 Real World Application with x265

In order to demonstrate the feasibility of achieving target constant quality using the proposed framework, we make use of the real-world encoder x265 and incorporate the constant quality control algorithm in the video encoder utilizing eGQP model. As discussed in the introduction section, one of the usage of GQP accurate modelling is to enable video encoders to choose appropriate encoding parameters so that the target constant video quality in terms of human visual perception is achieved. In Algorithm.1, the first input is the source video v with a total of M frames, where the i-th frame is represented as f_i. The other input is the target quality level represented as q in any video quality metric. S is used for representing the number of samples used for reconstructing the GQP curves, which is equal to 3 in this experiment. In the end, an encoded video with constant quality q is expected to be the output of the algorithm. 10 video clips of 6 frames in length from the test video dataset are selected for the experiment and demonstration. The encoding presets for the x265 encoder are described in Table.4.7 where I, P and B frames are all included in the final encoded video. The encoding order of the 6 frames of each video sequence is IBBBBP. The purpose of including I, P and B frames types in the encoded video is to show the generalization ability of the proposed framework for real world applications on different frame types. After sampling and analyzing the quality-QP data pair of each frame, the final decision for each frame is written in the QPFILE as described in the encoding command so that the final encoded video with the target constant quality is obtained. Fig.4.6 shows the

screen cuts for the 10 video clips, which covers a range of contents from humans, animals to natural scenes etc.. We make use of 3 basis/samples in the GQP function reconstruction part, which is represented as S in the algorithm. The target quality levels selected are 93, 96 and 99 in terms of the scaled SSIM, which in the range of $[0, 100]$. The selected quality levels cover the range of low, medium and high quality in terms of SSIM.

input : Source video v with M frames represented as f; Target quality level q
output: Encoded video in the target constant quality level

for $i \leftarrow 1$ **to** M **do**
  for $k \leftarrow 1$ **to** S **do**
    $f_{i,k} \leftarrow$ Encode f_i with encoder at x_k;
    $z_{i,k} \leftarrow$ `VQA(`$f_{i,k}$`)`;
  end
  Fit the quality-parameter (QP) function Q_i from $\{(x_{i,k}, z_{i,k})\}_{k=1}^{S}$;
  Select $x_{i,j} \leftarrow \min |z_{i,j} - q|$;
end

Algorithm 1: General Framework for Video Constant Quality Control

Table 4.7: Encoder Configurations

x265	x265 –input INPUT –keyint 50 –min-keyint 50 –no-open-gop –no-scenecut –b-adapt 0 –ipratio 1.0 –pbratio 1.0 –output OUTPUT –qpfile QPFILE

The frame level quality variation of the algorithm is shown in Fig.4.7. It can be seen that with only three samples for each frame, the target quality can be achieved with little variation for each frame of the video sequence. The per-frame quality variation curves demonstrate the accurate guidance of Algorithm.1 in achieving video quality on different quality levels.

The proposed approach provides a replacement of the current CRF method which doesn't consider the per-content quality consistency and therefore leads to significantly varying quality levels for the same CRF value across different source video content. In terms

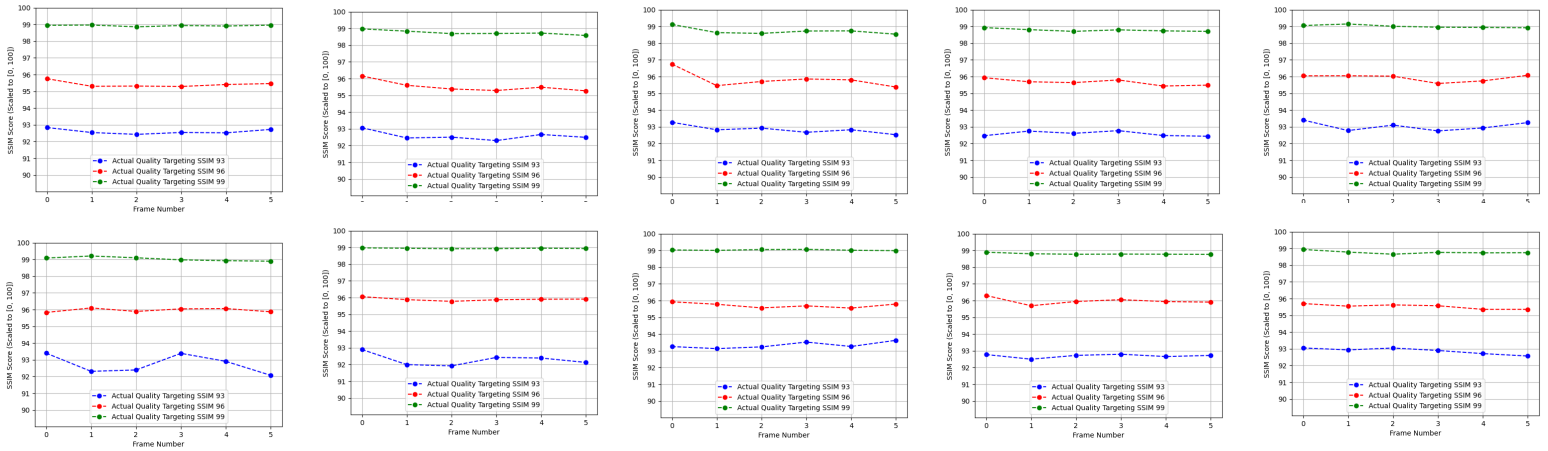


Figure 4.7: Frame Level Quality Variation Plot for Corresponding Source Contents

of time complexity, it should be noted that in the current implementation the sampling process of three samples per-frame is sequential, which is time consuming in terms of computational complexity. The algorithm can be easily adapted to concurrent sampling algorithm which takes much less time.

4.6 E2E Image Compression Quality Control with eGQP

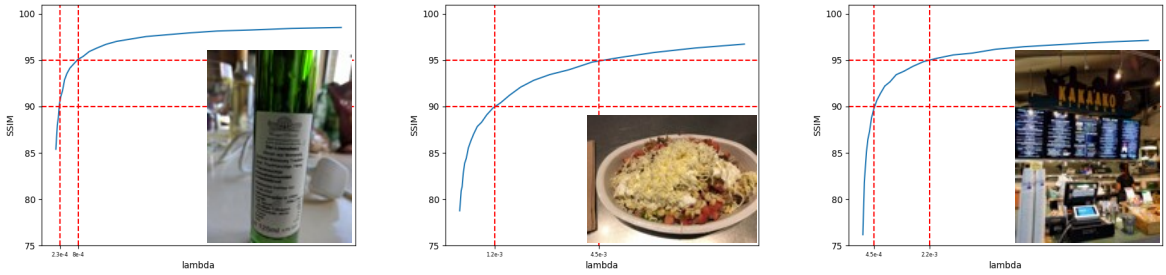
One factor that hinders the broad application of E2E learning based image compression method is the lack of accurate quality control mechanism to guide the image encoder to select the appropriate model for a specific λ value. Without knowledge of λ-quality modeling, aligned quality levels across encoded images is not feasible since the source image complexity varies across different content.

The eGQP framework not only helps video encoders to select the best set of parameters for a target constant quality level measured by human perception driven video quality metrics such as SSIM and VMAF, but it can also help recent learning based E2E image

compression method to select the controlling parameter for a target quality level. In this section, we will make use of the E2E method [88] for a demonstration. Firstly we will discuss the experiment setup and the general framework of how eGQP works in this task. Then we will show the effectiveness of eGQP method in selecting the hyper-parameter λ optimized for the neural network model.

4.6.1 Experiment Setup and eGQP Framework

Based on the description of the E2E model by Ballé et al.[88], different levels of rate-distortion trade-offs are achieved by training the neural network model using different λ values of the final optimization function. Therefore, firstly in the experiment, we train 50 models with 50 λ values evenly separated in the range of $[0.0001, 1]$ in the log scale. The database we use for training the models is CVPR image compression competition database [108]. 80% of the database is used for training E2E models and the same split is used for training the eigen basis for the eGQP framework. The rest 20% of the image database is used for testing the Algorithm.2 described in Fig.2. Secondly, by using the optimized models obtained from the first step, a quality-λ database is constructed, where the quality metric is SSIM and the encoding parameter is λ. In this experiment, the relationship between λ and quality is monotonic and meets the monotonic requirement of eGQP framework.



(a) (b) (c)

Figure 4.8: Samples of GQP curves for different image content compressed by E2E models.

In Algorithm.2, we elaborate the framework of selecting the appropriate λ optimized model for target quality q, which incorporates the proposed eGQP framework. In the algorithm, the samples are obtained through pre-determined optimized models according to the information theoretic uncertainty sampling introduced in [9]. After each sampling step, the image quality score is calculated using the selected IQA method, which is SSIM in this experiment. Finally, the quality-λ function is reconstructed using the eGQP framework and the appropriate λ that compress the source image to the target quality are selected. In Fig.4.8, we show three sample quality-λ curves with three source images. By marking the quality level 90 and 95, their corresponding λ values vary drastically, which proves the necessity of selecting different λ values across different contents to achieve the target quality. It should be noted that in Algorithm.2, the sampling process is sequential, while in real world applications, the sampling process can be easily adapted to a concurrent process, which would be much faster due to the independence relationship of each sampling process.

input : Source image I; Target quality level q
output: Encoded image in the target quality level

for $k \leftarrow 1$ **to** S **do**
    $f_k \leftarrow$ Encode I with E2E model at λ_k;
    $z_k \leftarrow$ `IQA(`f_k`)`;
end
Fit the quality-λ function Q from $\{(\lambda_k, z_k)\}_{k=1}^{S}$ with eGQP modelling;
Select $\lambda_j \leftarrow \min |z_j - q|$;

Algorithm 2: Control Parameter Selection for E2E Image Compression Quality Control

4.6.2 Performance of eGQP Framework on E2E Compression Quality Control

In Table.4.8, we show the average accuracy of quality control algorithm using RMSE and l_∞ as the two metrics with different number of sample/basis. It can be seen from the results that the average accuracy is reasonably good when the number of bases equals to

5, beyond which the quality improvement may be considered negligible [75]. Moreover, for the worst case scenario of RMSE and l_∞ metric, the quality level selection framework works as expected and successfully maintains the quality at an acceptable error range.

Table 4.8: Mean and worst performance of eGQP when the number of basis vectors is equal to the number of samples

N/S	RMSE		l^∞ error	
	Mean	Worst	Mean	Worst
1	2.81	16.68	7.01	33.73
3	0.39	1.79	1.45	9.42
5	0.23	0.77	0.80	5.24
8	0.15	0.52	0.70	2.56
10	0.14	0.45	0.61	2.03

Moreover, we conduct the experiment using Algorithm.2 on the testing split of the image GQP database with target SSIM quality levels 93, 96 and 99. In Table.4.9, the RMSE between the target quality and the actual quality across all the testing images are shown. As the number of samples/basis increases, the accuracy of the actual quality against target quality increases. When the number of basis samples/basis reaches 3, the quality RMSE, which is around 0.4, reaches a negligible level [75].

Table 4.9: RMSE of quality control algorithm at different quality levels

N/S	SSIM		
	93	96	99
1	3.05	1.55	0.36
3	0.47	0.43	0.26
5	0.31	0.17	0.12
8	0.26	0.16	0.07
10	0.24	0.15	0.06

4.7 Conclusion

Accurate, robust and computationally efficient models for video/image encoder control has always been the ultimate goal for researchers working in the area of data compression. With the advancement of machine learning methodologies, many encoder control models based on neural networks have been proposed in recent years [109, 110, 111, 112]. However, given the current robustness and computational complexity requirements of modern video/image encoders, most neural network based models are still steps away from real world applications. In this work, we provide a mathematical framework to analyze the relationship between image/video quality and encoder parameters and propose a data driven eGQP framework which is capable of reconstructing GQP functions accurately. The eGQP model connects the perceptual quality with encoding parameter in a more concrete and precise way than the currently widely used models with fixed empirical function forms. Unlike bitrate control, quality control has been paid less attention in the video compression research area due to the lack of concrete modelling between quality and encoding parameters. Targeting at a better perceptual quality control for both video encoders such as x265 and image encoders such as recently proposed E2E methods, the proposed eGQP framework makes it possible for encoders to choose the appropriate encoding parameters so that image/video can achieve desired perceptual quality with only a few samples that may be obtained concurrently. The effectiveness of eGQP framework is shown through two experiments inspired by practical scenarios, one for x265 video encoder and the other for E2E image encoder. Highly promising results are obtained in both experiments.

Chapter 5

Source Content Characterization and Selection by RD Domain Submodularity

The space of visual image content is extremely high dimensional if we regard the value of each pixel as a single dimension. With such a high dimensional data space at hand, it is impossible to exhaustively scrutinize each content for visual processing related tasks such as compression and visual quality assessment. The practical approach is to select several representative contents from the data space so that models for specific vision tasks may be trained, validated and tested. Therefore, selecting the representative visual source content is crucial. This chapter focuses on solving the problem of representative source content selection for image compression and related quality assessment applications. Nowadays, the source contents for compression and quality assessment applications are still hand-picked by researchers based on a few features such as content type, Colourfulness (CF), Spatial Information (SI), which are intuitively sensible but purely empirical measures with no justification on their diversity or representativeness of the visual image space. With the proposed visual content characterization method in the thesis, encoding Rate-Distortion (RD) analysis, which is capable of describing visual content using image encoders as analyzers, makes it possible to conduct source image selection for compression and related

quality assessment tasks since RD statistic is the direct result of lossy image compression. With the compression related characteristic at hand, we frame the source content selection problem as a subset selection process, where the representativeness can be modelled as a submodular set function. To the best of our knowledge, our work is the first effort to systematically address the source content selection problem for image compression applications. Through the experiments on deep neural network (DNN) based End-to-End (E2E) image compression method, we show that the source contents selected by our method are more representative as evidenced by their capability in boosting the performance of learning based image compression models.

5.1 Background

Visual data such as image contains lifelike information. Modern digital image contains millions of pixels, each at least represented by a scalar value (often quantized to a 8-bit unsigned integer). Since each pixel of digital image can be treated as one dimension of the data space, the dimensionality of the image space is extremely large. With the countless possible pixel combinations, it is a general belief that natural images watched by human visual system only span an extremely tiny cluster within the image space [113]. However, the number of contents in such a tiny cluster is still so large that one viewer cannot watch a small fraction of them in his/her lifetime. According to the online survey of [114], up to three billion images are shared online everyday in the year of 2022 and the number of photos taken is expected to reach more than one trillion by 2022. In contrast to the almost countless number of compressed digital images, the number of contents in compression quality related dataset only takes a tiny fraction which would be only a few thousand at most. What is more, since the image quality dataset usually requires subjective quality ratings, the time budget and the number of human participants limit the number of source contents that can be selected. Therefore, representative source contents need to be selected for a compression or related quality assessment dataset.

Even though representative source content selection is becoming increasingly important for compression applications given the proliferation of online image contents nowadays, the

number of studies on this topic is still limited. There may be two reasons behind the limited attention on it. The first one is the lack of reliable and commonly accepted compression quality related visual content characteristics. The second is the lack of reliable source content selection framework based on the visual content characteristics.

The first factor, i.e., visual content characterization, is still far from perfect. For example, as one of the responsibility of the video quality expert group (VQEG) is to produce visual content databases for compression related tasks [115], the VQEG group has proposed several visual quality databases for compression applications such as encoder performance comparison and quality metric validation. However, it is a common practice for them to select source contents based purely on artificially pre-determined content types such as natural scene, human, indoor, and outdoor etc. for selecting representative source contents. Since the compression task is believed to be a low-level vision task which involves the manipulations of visual frequencies and color etc., content types based on image semantic meaning can hardly be a reliable visual characteristic for compression applications. As shown in Fig.5.1, the Waterloo exploration image quality database [116] contains different content types ranging from human, animal, to landscape etc. Likewise, the HEVC video compression benchmark dataset contains video contents ranging from natural scenes, humans to animals etc [117]. Some intuitive and heuristic measures have been proposed with the belief that they are suitable for compression applications. In [3], SI using filter based method is proposed. The paper assumes that edge information plays a key role in determining the complexity of image contents and therefore affecting the level of difficulties for image compression. In ITU standard [4], the aforementioned concept is further modified and utilized to characterize video source content SI in order to build a representative video compression quality database. Many image and video compression datasets utilized SI on demonstrating the representativeness of their proposed dataset. [68, 44, 18] Moreover, with the increasingly focus on the High Dynamic Range (HDR) content, CF has often been used as another metric to characterize the image content from color perspective. The CF is based on the differences among red, green and blue channels [22]. In image compression quality dataset Waterloo Exploration II [2], the two aforementioned metrics SI and CF are used as proxies for demonstrating the good coverage and representativeness of the selected source contents. The scatter plots with convex hull are drawn for the proposed database

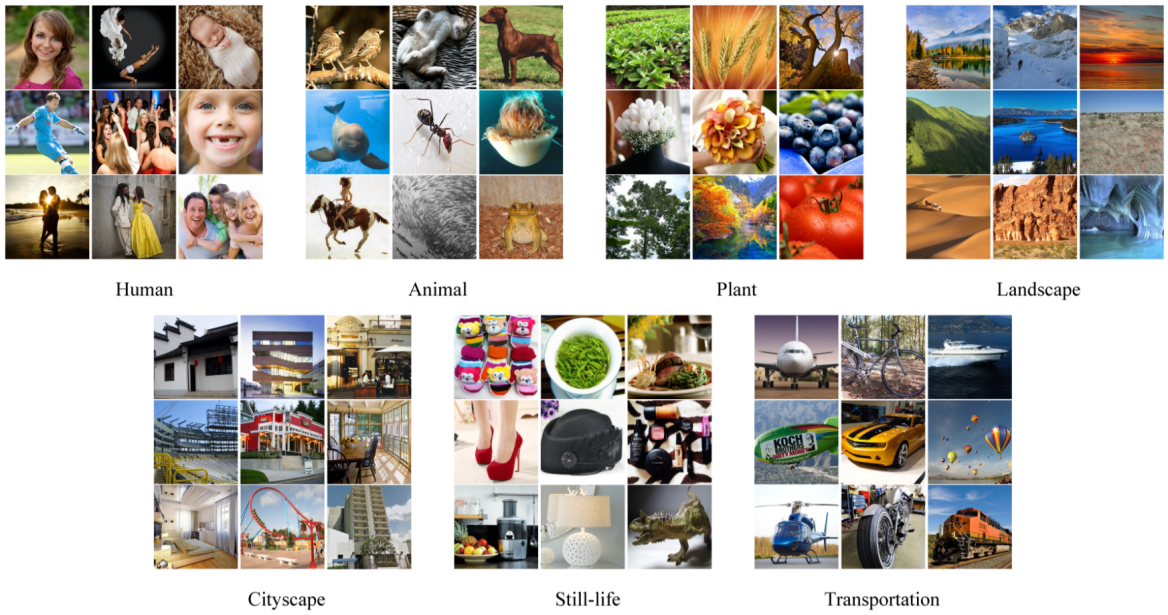


Figure 5.1: Waterloo exploration dataset content types

together with other 9 popular subjectively rated image quality databases, which are shown in Fig.5.2. However, there are several problems with such purely heuristic visual traits. Firstly, there is a lack of solid proof of the relationship between such visual traits and the compressibility of the image source content. Secondly, since image contents are complex and the research on Human Visual System (HVS) characteristics is still in progress, the practice of only counting on those limited number of visual traits is unreliable for describing images for compression applications. Lastly, the effectiveness of these over-simplified measures on capturing source content characteristics for compression applications is questionable. Notice that the original designs of such visual traits often take computational convenience as a major consideration.

For the second factor, the source content selection, Fig.5.3 shows a typical workflow of database construction for quality related tasks. Firstly, a large number of source contents are collected from the Internet or real-world photographs as the base data pool, for which the number could be large up to several millions. Then after keeping the pristine contents during the first round of visually filtering, a subset of the contents need to be selected based

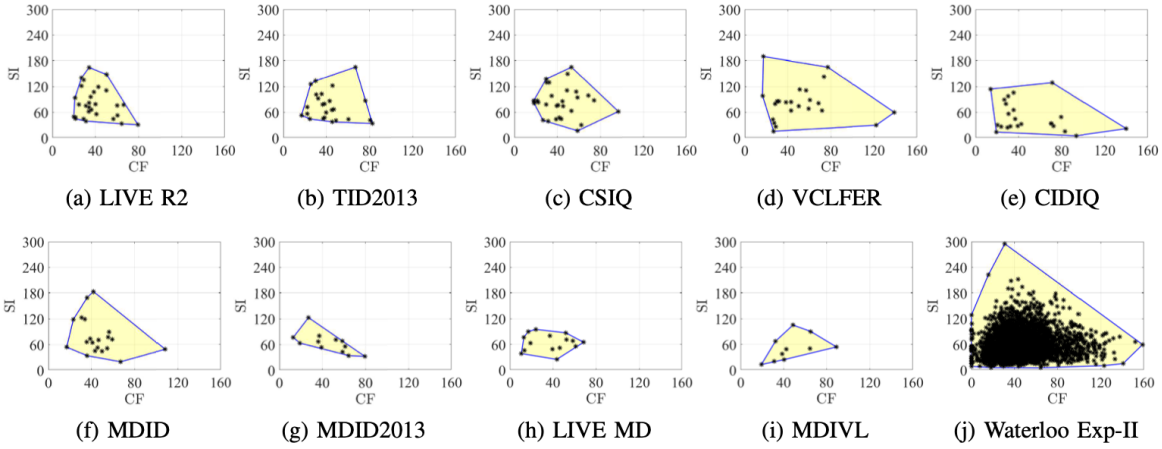


Figure 5.2: Scatter plots with convex hull for SI and CF in original Waterloo Exploration II image quality database paper. [2]

on the target task, for example, image compression or image quality assessment. Since the computational resources and the human participants for subjective experiment are limited, the contents in the subset can only be a fraction of the original image data pool. Therefore, a natural question is how to select representative source contents that is optimal for the subsequent tasks? It is a common practice in both image and video research area to count on "expert experience" to judge if the content is representative enough. It is obvious that this approach is not ideal. Firstly and most obviously, researchers' time is precious. One needs to watch every content in the data pool before determining which source contents are representative, not to mention the need of memorizing all visually checked contents. Given the reality of the explosive number of online images nowadays, the representativeness judgement through "expert experience" is an impossible mission. Secondly, humans have biases. Since humans have different visual preference and taste, the source contents subset selected by different image experts would differ from each other, which would be an inevitable factor that results in instability in the subsequent research tasks. Therefore, "expert experience" is not a scientific approach for image source content selection. Given that "expert experience" is unreliable, why not use random selection? Random selection has its drawbacks as well, especially when the subset only takes a tiny fraction of the original data pool. As mentioned in the description of Fig.5.3, the tight resources available in quality

related tasks cast a limitation on the number of source contents that can be selected. If the subset is randomly selected, the variations would be large and therefore wildly affect the outcome and reproducibility of the subsequent tasks. The second drawback is the possible lack of representativeness. Even though in theory random selection will asymptotically approximate the real distribution of the data, when the number of the selected samples is very limited, the samples are often incapable of offering a fair representation of the real data, especially, for example, when the distribution is heavy-tailed. Moreover, the best subset may be application dependent. For example, the compression and related quality assessment are the target tasks in mind, then it would be meaningful to select the source contents based on source content compression related characteristics.

Just like other research areas, database is the fundamental building block for compression quality research. Due to the limitations such as computing resources, time, and human participation, the number of source contents has to be limited. The lack of compression related visual characterization measures and the absence of systematic scientific source content selection approach lead to the limited progress made in the research area. Therefore, in this work, we aim to propose a mathematically sound and practically tractable method, for source content selection based on RD domain submodular optimization.

5.2 Motivations

5.2.1 Encoding RD Analysis

In order to address the first problem of lacking task related visual content characteristics, we propose to use encoding RD analysis.

In this thesis, we have demonstrated the usefulness of encoding RD analysis on precise encoder quality control, which demonstrates that it is capable of capturing visual content characteristics from compression perspective. Moreover, as described in the work of [1, 9], many multimedia applications require RD functions to characterize source signals and maximize Quality of Experience (QoE). Examples of applications that explicitly use RD measurements include but are not limited to codec evaluation [118], rate-distortion

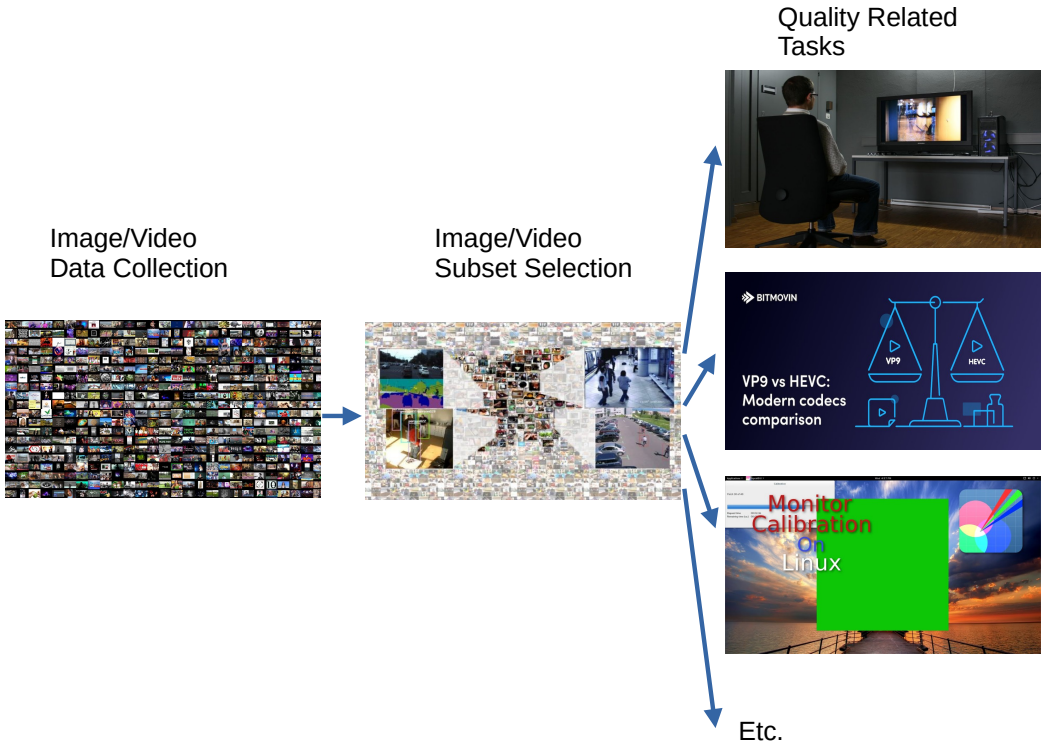


Figure 5.3: Quality database construction process

optimization [7], video quality assessment (VQA) [119], encoding representation recommendation [120, 121], and QoE optimization of streaming videos [122, 123]. The RD statistic is a natural fit for visual content characterization of compression applications. Firstly it is obtained from visual compressors, which is a system that aiming for describe the source signal in an efficient way. As shown in Fig.5.4, different content has drastically different RD behaviours. Secondly, unlike other established visual content characteristics that is only represented by a single number, the RD statistic is a progressive descriptor of source content in that the encoded signal's quality or distortion increases with the increment of bits used to describe them. In practice, it is less likely to find two source contents sharing the same RD curve while it is common to see two different content with the same single number descriptor value, as shown in Fig.5.5. The two images share the same SI value of 68 while their RD curves behave differently. The two curves are obtained by lossy encoding the two images into 25 different quality levels using JPEG image encoder. Their SI values are evaluated using the original source images shown in the RD plots.

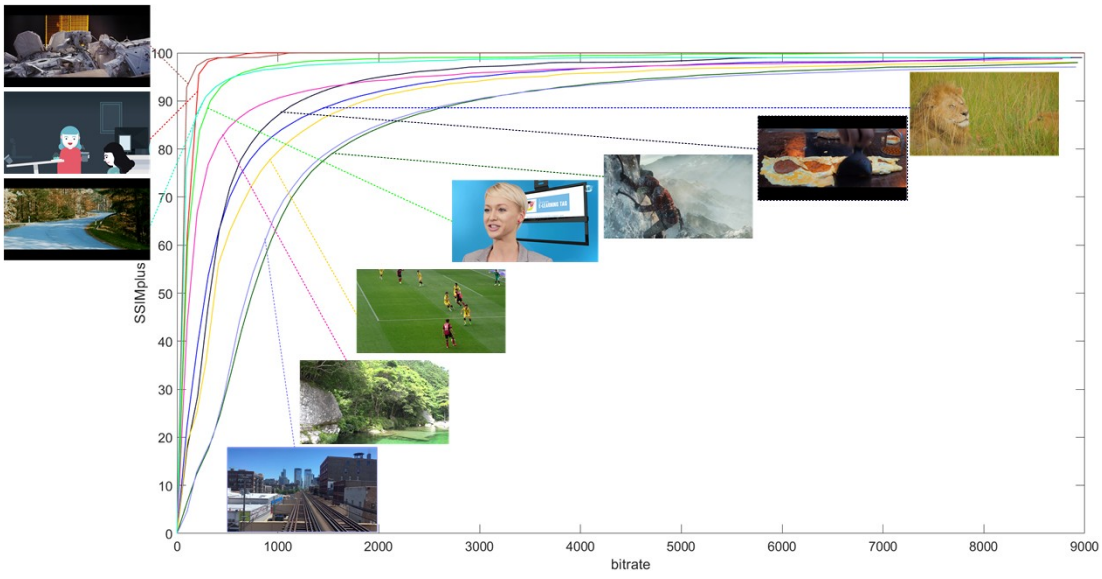
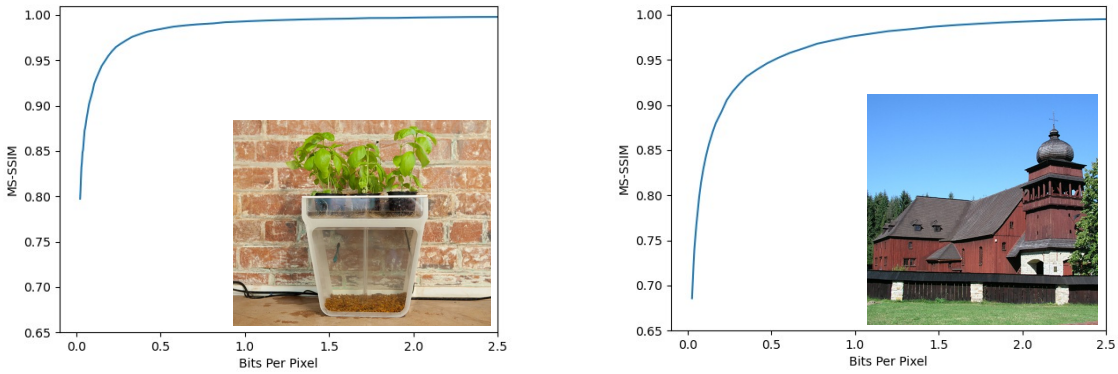


Figure 5.4: Different contents behave differently in RD domain



(a) (b)

Figure 5.5: Samples of RD curves for different image content with the same SI

5.2.2 Submodularity

Submodularity is defined in the context of set function whose domain is a family of subsets of a given set, called ground set, that usually takes its value as a real number. The set functions are discrete functions that find their application in combinitorial problems. By assigning values to the subset, it is possible to mathematically find the optimal set given some conditions.

There is a special type of set function called submodular set function, which has the diminishing return property. Mathematically speaking, if Ω represents the ground set, and the set function $f : 2^{\Omega} \rightarrow \mathbb{R}$ is submodular, then we have:

For every $X, Y \subseteq \Omega$ with $X \subseteq Y$ and every $x \in \Omega \setminus Y$, where 2^{Ω} denotes the power set of Ω:

$$f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y) \tag{5.1}$$

The definition above illustrate the "diminishing return" property that is common for many real world problems. For example, for the source content selection problem in this

study, the representativeness naturally has the "diminishing return" property. Assume representativeness is denoted by set function R with a value range of $[0, 1]$, the value of R for the ground set would be maximum when all contents are selected and minimum when there is no content selected. Since in each selecting step, the most representative source content can be selected from the unselected ones, each following selected contents' value R would be less than their previous one until all contents in the ground set are selected. In Fig.5.6, the diminishing increment of R for the optimal selection procedure is demonstrated in a continuous way.

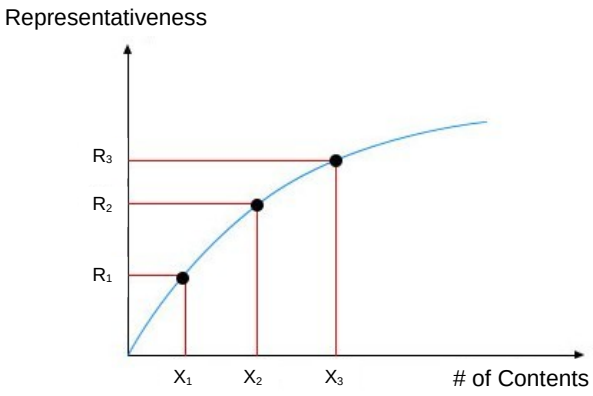


Figure 5.6: Diminishing Return Property of Representativeness

In this work, we use the facility location function for measuring the selected contents' representativeness in RD domain. Facility-Location function [124] attempts to model representation, as it tries to find a representative subset of items from the ground set, which are denoted as X and V, respectively in Eq.5.2. The function ϕ can be a measure of similarity. In this study, we use the inverse of Euclidean distance as the similarity measure.

$$R(X, V) = \sum_{v \in V} \max_{x \in X} \phi(x, v) \tag{5.2}$$

The facility location function is a submodular function. For this function, every element $v \in V$ must have a representative within the set V and the representative for each $v \in V$ is chosen to be the element $x \in X$ most similar to v. This function is also a form of dispersion or diversity function because, in order to maximize it, every element $v \in V$ must have some element similar to it in X. The overall score is then the sum of the similarity between each element $v \in V$ and v's representative. [11]

Thanks to the wide availability of submodular optimization toolbox such as Apricot [125], the source content selection problem can be conveniently modeled using the Eq.5.2 and optimally solved.

5.3 Source Content Selection Procedure

Fig.5.7 shows the framework for the proposed source content selection procedure. The following is a description of each step:

1. The selection procedure starts with the collection of source pristine images. In our work, we assume that the source images have gone through a visually check step for their pristineness.

2. In the second step, RD information is collected from the ground set images using an image encoder. Since modern image encoders usually provide users with the flexibility to adjust the bitrate or quality for the encoded images in lossy compression mode, the RD analysis step can be easily conducted. It should be noted that any lossy image compressor can be used in this step for RD analysis. In the experiment section we will prove the choice of lossy image compressor has limited effect on the final compression applications.

3. In the third step, submodular optimization is conducted using the facility location function, which takes the RD statistics as the input. The subset is obtained after the submodular optimization. The Apricot submodular optimization toolbox [125] is utilized in this step.

4. Lastly, the database are constructed based on the selected subset for compression applications.

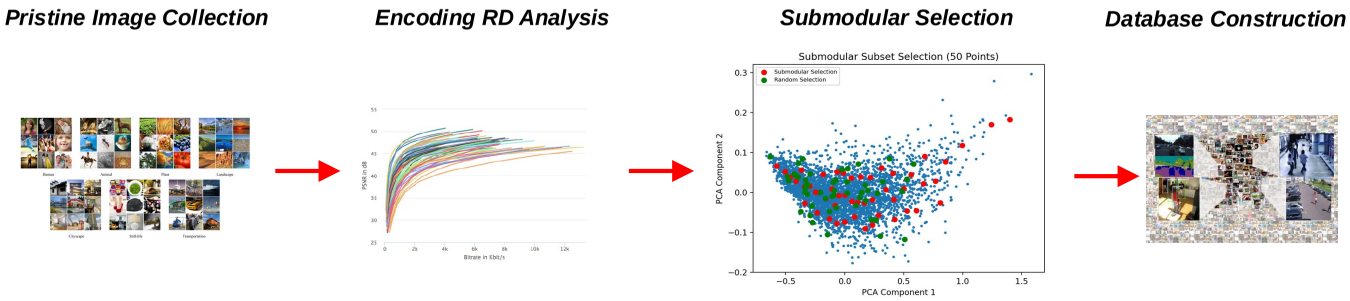


Figure 5.7: Proposed framework for source content selection

5.4 Experiment

In order to verify that the proposed framework effectively captures the representative contents, we conduct an experiment using an E2E learning based image encoder [12]. Since the E2E image compression model is a learning based model, the representativeness of training database predominantly affects its compression performance. Therefore, training learning-based image encoder offers an ideal testbed for approve or disapprove the proposed framework.

5.4.1 Image Encoders for Encoding RD Analysis

We select five widely used lossy image encoders, JPEG [126], JPEG2000 [127], AVIF [128], HEIF [129], and WEBP [130] for encoding RD analysis. Below is a brief description of the five encoders:

1. JPEG image encoder is proposed by Joint Photographic Experts Group (JPEG), which is widely adopted and supported by most image platforms. The JPEG encoder is designed using the discrete cosine transform.

2. JPEG2000 is proposed by the JPEG group using the discrete wavelet transform, with the design goal of more functionality and better RD performance compared with JPEG.

3. High Efficiency Image File Format (HEIF) is a container format for storing individual digital images and image sequences. It supports high quality image encoding using High Efficiency Video Coding (HEVC) intra encoding mode.

4. AV1 Image File Format (AVIF) is an image file format specification for storing images or image sequences compressed with the recent video coding standard AOMedia Video 1 (AV1) using its intra coding tool.

5. WebP is developed by Google in 2010 as an open-source standard for web-based true color graphic images. It is claimed to maintain JPEG image quality with a smaller file size.

5.4.2 RD Domain Selection and Training Procedure

The Waterloo Exploration Database [116] is used in this work. The image dataset contains 4,744 pristine natural images and 94,880 distorted images created from them. The database is split into 80-20 training-testing partitions. For our work, we make use of the 80 percent training part images as the ground set for image training subset selection. The evaluation is conducted using the testing partition.

Firstly, all images from the training partition of Waterloo exploration database [116] are lossy compressed using the five image encoders into multiple compression levels, from which the RD domain statistics for each content are obtained. The rate measured in bits per pixel and quality measured in SSIM [131] are recorded for each content per encoder. Secondly, 5%, 10%, 20%, 30%, and 40% training images are selected from the training partition using the proposed submodular method based on the RD domain statistics. The selected images are used for training the E2E image encoder. As a comparison set, another sets of randomly selected 5%, 10%, 20%, 30%, and 40% images are generated from the training partition. Lastly, multiple E2E image compression models are trained separately using the source contents selected by the proposed method and randomly selected source contents. We perform random selection five times for each subset percentage in order to prove the proposed method is constantly better than random selection. The E2E image encoders are trained using the same setting according to the original paper of E2E image compression method [12]. For convenience, the encoding models trained using submodular selected images are called ss_models and the encoding models trained using randomly selected images are called rs_models.

5.4.3 Encoding Performance Comparison

The performance is compared between the ss_models and rd_models on the testing partition. In order to quantify the coding gain or rate savings achieved by the ss_models compared against rs_models, we choose to use the Bjøntegaard Delta (BD) method [57, 58], which is widely used in the visual compression field to evaluate the relative coding efficiency of one codec against a reference codec [132] over a range of quality-bit rate data

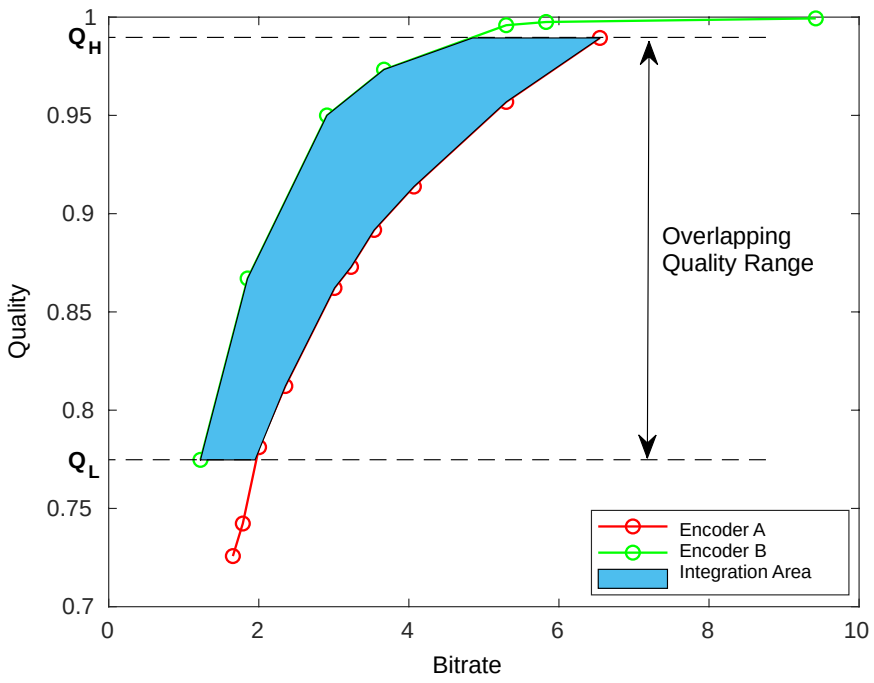


Figure 5.8: Sample RD curve comparison and BD-rate computation.

points. Given two RD curves produced by two encoding models, ss_models and rd_models, we compute the BD-rate metric, which estimates the average bit rate savings for the same quality (in terms of SSIM). The bit rate saving for a given level of quality is calculated as

$$\Delta R(Q) = \frac{R_B(Q) - R_A(Q)}{R_A(Q)} \tag{5.3}$$

where $R_A(Q)$ and $R_B(Q)$ are the bitrates for quality level Q on the reference and test RD curves, respectively. Since the logarithmic scale $r = \log R$ is used in the BD model on the bit rate axis, the bitrate saving can be expressed as

$$\Delta R(Q) = 10^{r_B(Q) - r_A(Q)} - 1 \tag{5.4}$$

Considering both the actual RD points and the fitted RD curves $\hat{r}(Q)$, the BD-rate can be approximated by

$$\Delta R_{Overall} \approx 10^{\frac{1}{Q_H - Q_L} \int_{Q_L}^{Q_H} [\hat{r}_B(Q) - \hat{r}_A(Q)] dQ} - 1 \tag{5.5}$$

where Q_H is the maximum of the minimum quality that the two curves could reach, and Q_L is the minimum of the maximum quality that the two curves could reach. The region of integration is exemplified as the blue region in Fig. 5.8.

In Fig. 2.12, we showed a bar plot for rate-savings of the ss_model compared against rs_model using different RD analysis encoders and under different training subset fractions. We have several observations:

1. The ss_models are better than rs_models in terms of average rate-saving for the testing partition. The numbers on the bar plot show the average rate-saving percentage while the "T" bars show the standard deviation of the rate-savings. It should be mentioned that rate saving performance gain rarely goes beyond 5 percent without major architecture changes made to the encoder. Therefore, the 3 percent rate saving performance increase purely by changing training sets is a significant boost.

2. The ss_models trained on different image encoders are color-coded in Fig. 2.12. It can be seen that the proposed RD domain selection method is robust in terms of different encoders' RD characteristics, which further validate that encoding RD analysis is a useful visual content characterization method.

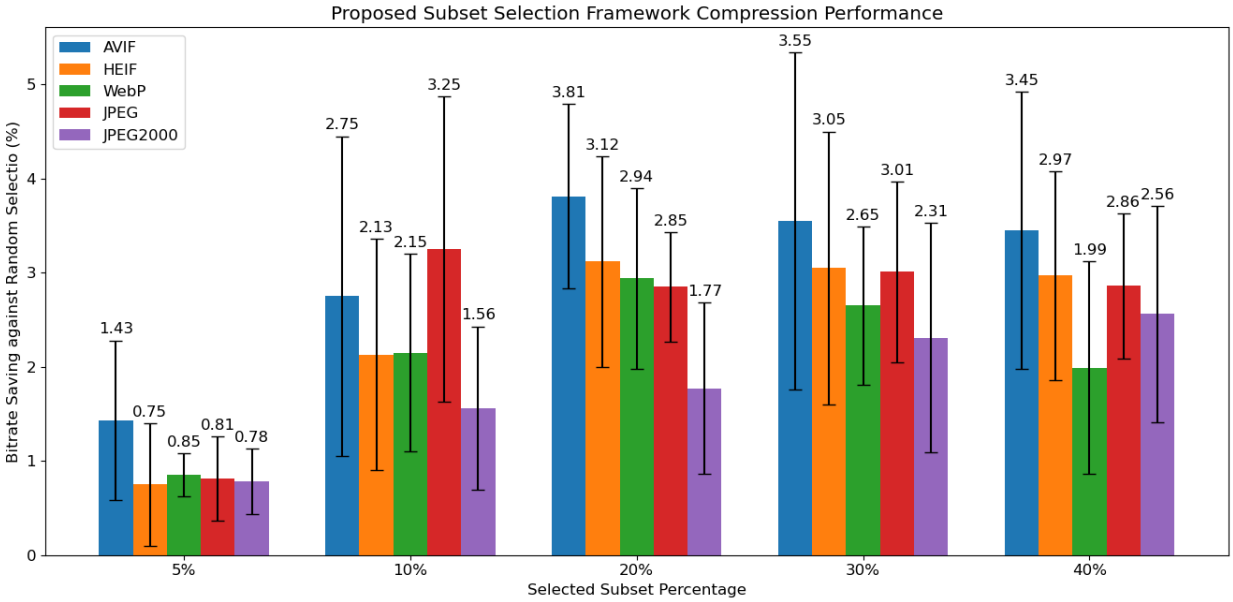


Figure 5.9: Performance comparison of submodular trained model using different encoder's encoding RD analysis against the random selection measured by rate-saving percentage, the standard deviation bars are obtained based on the comparison of selected subset against the five times random selection for each subset percentage per encoder.

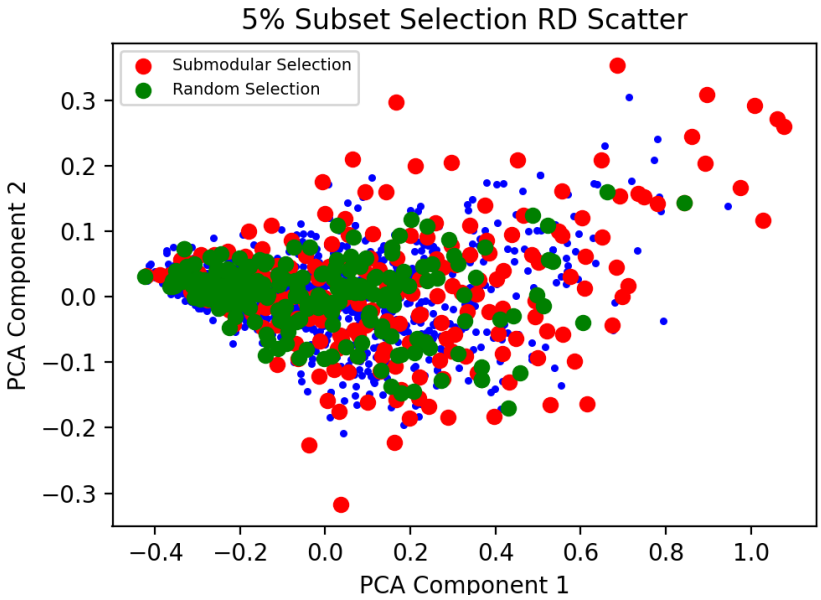


Figure 5.10: Submodular RD Domain Selection VS. Random Selection

3. The 5% subset rate-saving is the least compared to other cases. Since the 5% image subset only contains about 180 images for each encoders' RD domain selection, the training model is prone to over-fitting. Therefore, the rd-saving gain of ss_models is not as strong when compared with the cases of larger subsets.

In Fig.5.10 we show a RD domain scatter plots comparing the submodular selection and random selection. For the convenience of visualization, we showed the first two Principal Component Analysis (PCA) components on the RD curves of AVIF image encoder, which explain over 90% of the original data's variation energy. As can be seen in the figure, when compared to random selection, submodular optimization method selects more widely spread data uncovered by the randomly selected points, which leads to much better rd-saving boost in those under represented RD domain. For example, for the test image shown in Fig.5.11, the rate saving is over 10 percent. The example content's RD domain data point is marked as black in the scatter examples shown in Fig.5.11 and Fig.5.12. For both examples, as can be seen in the scatter plots on the right, there are not many randomly selected data points around the test image in the RD domain, but since there are points

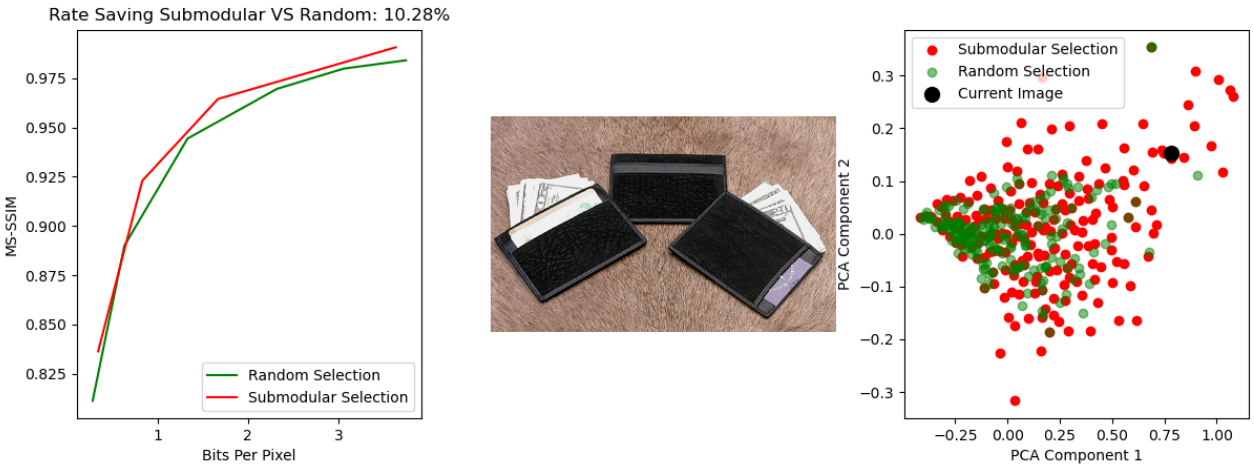


Figure 5.11: Sample Content for Submodular RD Domain Selection VS. Random Selection

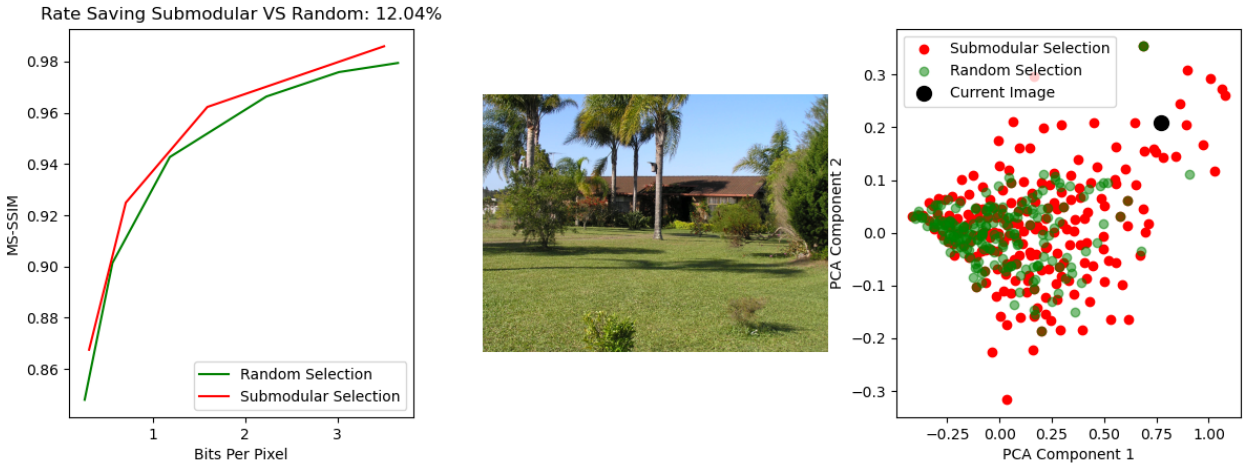


Figure 5.12: Sample Content for Submodular RD Domain Selection VS. Random Selection

around it selected by the proposed method, there is a rate saving of over 10%. Moreover, we can observe that even though there are fewer samples selected by the proposed method in the area most densely sampled by random selection, the ss_models rd-saving performance is still better on average. These observations suggest that the proposed method avoids over-sampling or duplicated sampling at the well represented regions in RD domain, and thus leave space for more samples in the under represented RD regions, where the most coding gain is obtained.

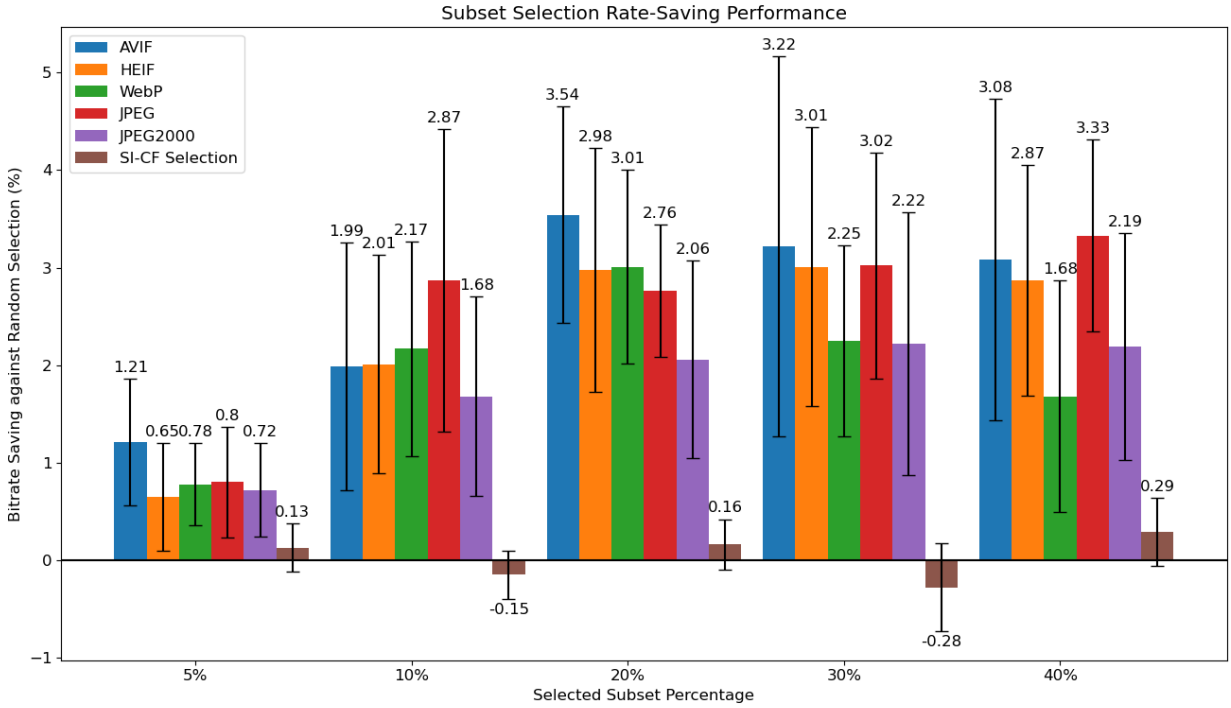


Figure 5.13: Performance comparison of submodular optimized model based on source images' SI and CF selection against the random selection measured by rate-saving percentage. The standard deviation bars are obtained based on the comparison of selected subset against the five trials of random selections for each subset percentage.

5.4.4 Visual Characteristics Comparison

In order to validate the effectiveness of RD characteristics against the traditional visual features widely used in the literature, we compare the features selected in RD domain with those of SI and CF. For fair comparison, the same submodular optimization is applied in both cases. Specifically, we perform a PCA analysis on the RD statistics and use the first two components to rerun the RD domain source content selection experiment. In order to make a controlled comparison, SI-CF domain selection follows exactly the same procedure as RD domain selection described in previous sections.

Fig.5.13 shows a bar plot for rate-savings of SI-CF domain selection compared to random selection, together with the new RD domain selection results for different image encoders. For convenience, the encoding models trained using submodular selected images based on SI and CF are called ss_sicf_models while encoding models trained using submodular selected images based on RD are called ss_rd_models. Based on the bar plot, it can be observed that the performance boost based on SI-CF is negligible compared to the RD domain selection. On average, the rate saving across the five subset percentages are only around 0.03 for SI-CF domain selections. Moreover, in order to verify the observation is statistically significant, we perform t-tests on the rate-savings of ss_sicf_models against ss_rd_models. The null hypothesis for all the t-tests is that the difference between rate-savings of SI-CF selection and RD selection is zero. Since there are five image encoders available, we conduct t-tests for each of them with the corresponding ss_sicf_model at the five different subset percentages. The p-values and t-statistics are shown in Table.5.1 and Table.5.2 respectively. It can be seen that the p-values are all below 0.05, and thus the null hypothesis is rejected, which means that the performance boost of using RD against traditional method SI-CF is statistically significant.

More insights may be obtained from Fig.5.14, which shows the SI-CF domain scatter plot that compares the submodular selection and random selection. Similar to the scatter plot of Fig.5.10 for RD domain selection, submodular optimization method selects more widely spread data uncovered by the randomly selected points suggesting that the submodular optimization process is working properly. An important and perhaps more interesting observation is that the SI and CF are positively correlated, with a significant

Table 5.1: T-Test Results (P-Values) for RD Domain Selection vs. SI-CF Domain Selection

Selection Percentage	P-Values for RD vs. SI-CF Selection				
	AVIF	HEIF	WebP	JPEG	JPEG2000
5%	0.012	0.020	0.001	0.012	0.003
10%	0.001	0.001	0.001	0.043	0.010
20%	0.002	0.001	0.001	0.001	0.001
30%	0.001	0.016	0.002	0.001	0.001
40%	0.003	0.001	0.001	0.001	0.021

Table 5.2: T-Test Results (T-Statistics) for RD Domain Selection vs. SI-CF Domain Selection

Selection Percentage	T-Statistics for RD vs. SI-CF Selection				
	AVIF	HEIF	WebP	JPEG	JPEG2000
5%	3.22	2.89	4.71	3.22	4.12
10%	4.95	5.06	5.35	2.41	3.36
20%	4.43	6.37	10.26	5.12	7.70
30%	5.87	3.04	5.49	6.19	8.12
40%	4.07	8.59	5.89	7.76	2.86

Pearson correlation coefficient of 0.68. By contrast, the RD domain features are linearly uncorrelated by design. This notable difference may explain why the previously widely used visual characteristics such as SI and CF are not efficient visual characteristics, since they do not capture independent information of the source contents. The correlation between SI and CF leads to the two-dimensional measures working in a somewhat redundant fashion while the eigen-RD modelling captures linearly independent principal components of RD statistics.

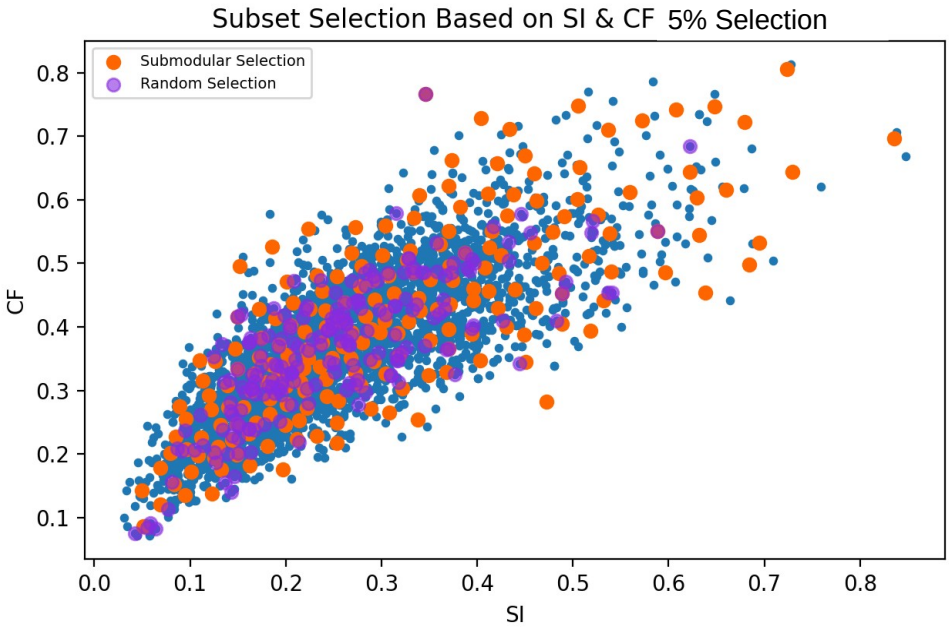


Figure 5.14: Submodular SI-CF Domain Selection VS. Random Selection

5.5 Conclusion

The selection of training and testing databases is a fundamental building block in the development of learning based visual applications. Specifically, for compression tasks, due to the limited time and computational resources, a dataset can only contain a very small set of visual content. On the other hand, due to the ambiguous relationship between the source content compressibility and widely recognized visual traits such as SI and CF, those traits have been found to be unreliable visual characterization methods in the process of database construction. Instead, "expert experience" is widely used when judging the representativeness for a specific content, which is subjective, inconsistent and again, unreliable.

In this work, we propose to use encoding RD analysis as a compression oriented visual

content characterization method. Combining it with the submodular optimization framework using a specific set function measuring the selected subset's representativeness, we make the first effort to solve the problem scientifically. Our proposed approach is easy to use and provides robust selection result when compared against random selection. The effectiveness of our method is demonstrated utilizing a learning based image compression method [12]. By comparing the trained model using source content selected based on the proposed method and random selection, we show our method is better in terms of average rate saving. Close observation suggests that the most significant savings come from the source content that is under represented by the random selection method, which in turn indicating a better representativeness of the source content selected by the proposed method.

The philosophy of maximizing representativeness provides novel insights on data selection for machine learning problems. Traditionally, matching distribution has been the golden standard of machine learning, especially for classification problems aiming to reduce mis-classification rates or regression problems targeting at lower mean regression errors. Both cases favor randomly selecting the training set or testing set. The assumption is that matching the source data distribution would lead to the optimal results on testing set since the testing and training sets share the same distribution. However, based on our investigation we found out the assumption may not always hold for some cases such as visual compression and quality related tasks. Our finding suggests the reason may be different source contents may carry different weights in the final evaluation because they may lead to different levels of performance boost or drop. There needs a balance between matching the distribution and maximizing the diversity of the source contents especially when the selected data is so limited compared to the available source data that can be selected from.

Potentially, matching source distribution and maximizing diversity could lead to drastically different source content selection. For example maximizing diversity would result in the selected contents to stay as far apart as possible in the feature space, the problem would be severe if the data distribution is long-tailed. On the other hand by matching the source distribution, the selected points would be clustered together sharing similar visual characteristics, which would negatively affect the learned encoders' performance since there is a lack of different visual contents with diverse characteristics, leaving uncovered content

types in the training data that the encoder performance is highly sensitive to. By contrast, the proposed method attempts to maximize representativeness and provides a way to balance between matching source distribution and maximizing diversity. As such, each selected content is a good representation of the data points surrounding it and each data point in the large data collection can have a representative point nearby in the selected subset.

Chapter 6

Conclusions and Future Perspectives

6.1 Conclusions

The thesis tackles the visual content characterization problem from Rate-Distortion (RD) analysis perspective. In the literature, for compression and related quality assessment tasks, a thorough investigation of the content characterization problem is lacking. Though several heuristic visual traits are used to describe visual characteristics, little theoretical or empirical justification has ever been provided and their reliability is found to be poor in real world applications. Inspired by the fact that RD analysis' content adaptation capability in compression related tasks and its extensive utilization for visual quality related tasks, we propose to use it as a measure of visual content characterization for compression and related applications. Moreover, we reviewed the process of source content selection procedure for several visual quality related databases. Due to the absence of a systematic approach on representative source content selection, purely empirical "expert experience" selection is still widely used nowadays. Scientifically sound methods are largely lacking in practice.

Through the 3840×2160 or 4096×2160 pixel resolution (4K) video encoder comparison project, the necessity of compression task oriented visual content characterization is revealed in the database construction step, in which source video contents are selected based on the content types. Moreover, because there is no suitable measure of source

contents' representativeness, selecting representative source contents has always been a subjective task where utilizing "expert-experience" is still a common practice. Furthermore, the encoder performance analysis using encoding RD analysis indicates that it is highly desirable to develop an accurate encoding parameter model against quality. In the subsequent chapters, all the aforementioned problems are addressed using encoding RD analysis as a visual content characterization method.

The effectiveness of RD analysis is demonstrated through an encoding RD analysis inspired visual signal compression quality control work. Inspired by the work of [1], the video encoders' quality parameter functions are reconstructed precisely through an eigen analysis approach with just a few samples, which outperforms the current widely used models that utilizes edge detection based visual trait Spatial Information (SI). With the proposed method, it is possible to control the encoded video's quality measured by Human Visual System (HVS) driven quality metric. Moreover, the proposed RD domain eigen-analysis based precise quality control framework makes it possible to control the encoded image's quality for image compression model driven by End-to-End (E2E) neural network. As such, we make the first effort to design encoder control mechanism for E2E image encoders without sacrificing compression efficiency.

Furthermore, the visual content characterization problem is addressed with the encoding RD analysis and submodular subset selection framework. Using the proposed characterization method, encoding RD analysis, which is capable of describing visual content using image encoders as analyzers, makes it possible to conduct source image selection for compression related tasks since RD statistic is the direct result of lossy image compression. With the compression related characteristic at hand, we frame the source content selection problem as a subset optimization process, where the representativeness can be modelled as a submodular set function. Our work is the first effort to systematically address the source content selection problem for image compression applications. Through the experiments on deep neural network (DNN) based E2E image compression method, we show that the source contents selected by our method are effective in boosting the performance of learning based compression model and demonstrates better representativeness of the selected content in RD domain.

6.2 Future Perspectives

The results in the thesis demonstrate the potential of applying encoding RD analysis on visual content characterization for compression and quality related applications. The thesis mainly focuses on applying RD characterization on two aspects, namely, precise quality control for visual encoders, and maximum-representativeness source content selection for database construction. Besides the two compression based applications, we believe the RD characterization philosophy has a vast potential for applications in other areas. Moreover, the representative source content selection for compression image database provides novel insights on data selection for quality related tasks, especially when the data points are abundant while not all of them can be utilized due to resource limitations in the labelling and training process. In this section, several possible future perspectives initiated from the ideas proposed in the thesis will be discussed.

6.2.1 RD Encoding Analysis for Quality Enhancement

One possible application of the RD encoding analysis is artifact detection for fast video content quality enhancement. Video quality enhancement can be divided into two separate sub-tasks, which are the artifact detection task and the quality enhancement task. The artifact detection task may utilize pattern recognition techniques to locate regions containing quality degradation either inside a single image or across multiple video frames. Current popular methods either utilize convolutional neural network (CNN) to take the whole image directly as input for analysis and enhancement as the end-to-end solution or rely on some ad-hoc features calculated based on the video frames for the detection task [133, 134, 135]. However, given the heavy computational burden of the CNN methods and the heuristic nature of the unjustified empirical features, the existing methods are both slow and unreliable to meet the need of time-critical tasks such as live video broadcasting. The RD characteristics, which are the direct results of encoding and have the proved capability on characterizing source contents, may be utilized for the artifact detection task in the video trans-coding step, which is a common and necessary video processing step in the video content delivery pipeline. Thanks to the block-based nature of modern visual

encoders, each Coding Tree Unit (CTU) can be encoded independently and carry corresponding RD statistics as its metadata. If RD statistics can be collected and recorded for each block or frame, which is almost effortless in terms of computational burden when compared to most CNN solutions, efficient detection methods based on RD characteristics can be achieved for time-critical video enhancement applications.

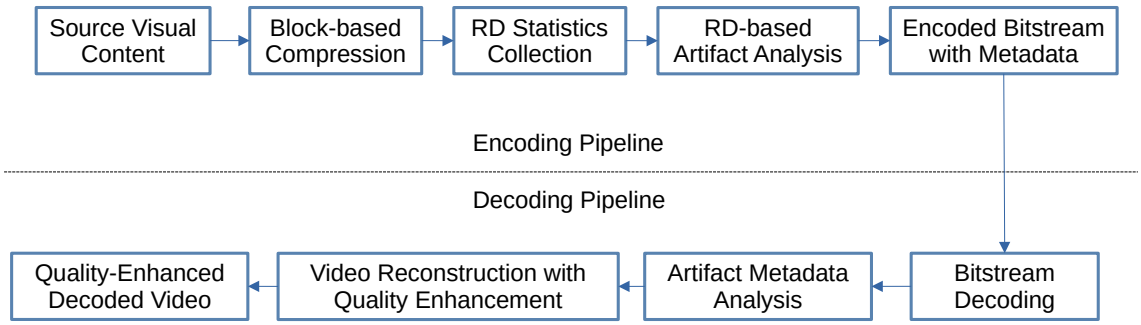


Figure 6.1: Video Quality Enhancement with RD Encoding Analysis Workflow

One possible workflow of applying RD analysis on encoding quality enhancement is shown in Fig.6.1. The two quality enhancement sub-tasks, artifact detection and quality enhancement, are deployed in the video encoding and decoding pipeline respectively. At the encoding side, RD statistics can be collected for each coding unit of the source visual content, which can be analyzed later in the RD-based artifact analysis module deployed on the encoder side. At the decoder side, the metadata containing artifact information will be decoded and analyzed so that a proper quality enhancement method can be chosen in the quality enhancement module deployed in the decoding pipeline. The direct integration of RD-based artifact analysis module and quality enhancement module in the visual content encoding-decoding pipeline would achieve efficient quality enhancement because RD information is easily accessible during the encoding and decoding process in the video delivery

pipeline.

In summary, RD characterization may provide a computational efficient and reliable solution when compared against the computationally expensive CNN based methods and the methods relying on unverified ad-hoc visual content features.

6.2.2 RD Encoding Analysis Extended to Other Signal Types

RD characterization philosophy can be further extend to the analysis of emerging visual content types such as 3D visual content, visual-reality content, and point-cloud visual content. The emerging visual content display technologies not only bring better immersive viewing experiences to consumers, but also create new challenges on building representative database because of limited content availability and high cost of content displaying and capturing devices. Thanks to the fact that most emerging visual content compression technologies are based on the established video encoders such as AOMedia Video 1 (AV1) or High Efficiency Video Coding (HEVC), the proposed RD analysis and modelling framework can be applied following the same procedure as discussed in the thesis without vast modifications. The RD analysis may shed light on the automated analysis on emerging visual content types so that researchers' burden of source content selection and generation would be alleviated because the RD analysis is totally based on the encoded contents' encoding statistics that are objective, automated, and can be obtained without human judgement involved.

With the success of applying RD analysis for visual content characterization, a more general application of using RD analysis for precise encoding quality control on more signal types is possible. Besides visual signals, the RD information of signals such as sound and haptic, may also be obtained by using their lossy compressors, which are widely accessible since the signals are intended to be transmitted under the limited bandwidth. Therefore, the generalized "compressor as analyzer" philosophy may be applied to a wide variety of lossy signal compressors for characterizing a broad range of signal modalities. In Fig.6.2, a general workflow of applying RD characterization is shown with several possible source signal types, such as audio, video, haptic, and 3D image. The RD statistics can be either utilized as features for downstream tasks such as pattern recognition or as signal database

source selection measures. On the other hand, proper modelling of the RD statistics may also create opportunities for developing precise signal quality control methods and quality enhancement methods for the lossy compressors of more signal types.

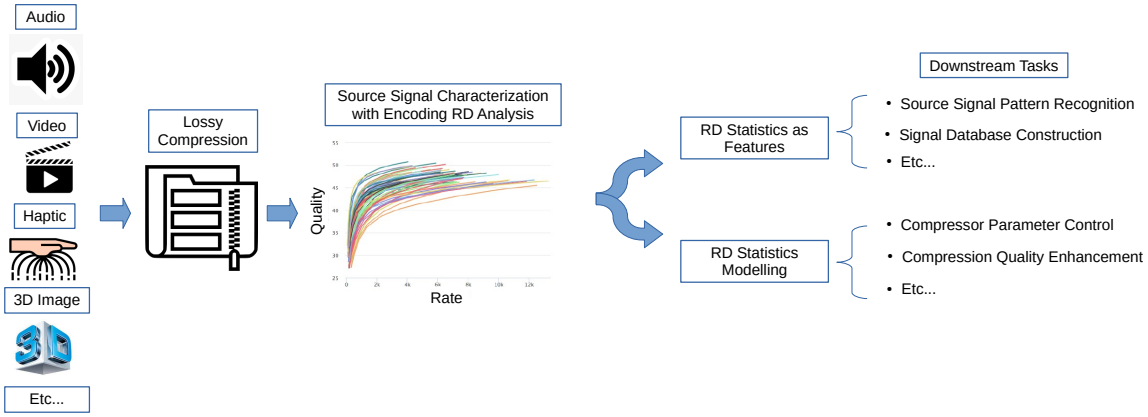


Figure 6.2: A General Workflow of Source Signal RD Characterization

It should be noted that different types of signals have different quality measurement methods. Cares should be taken for the selection of signal quality metrics due to different quality assessment procedures. In other words, the D term in RD analysis could be adapted to the specific signal types or the specific tasks. In summary, the RD characterization idea is highly extensible both in terms of applications and source signal types. Researchers may take the compression related RD statistics into consideration when working with data characterization related problems in the future.

6.2.3 Maximum-Representativeness Source Content Selection for Quality Related Tasks

The maximum-representativeness source content selection methodology has demonstrated its potential on boosting the performance of learning based image compressor. The methodology provides a practical solution on handling problems with abundant data while the

allowed training data for specific application is limited due to resource limitations such as high data labeling or computational cost.

Besides the compression tasks, another type of applications with great potentials are those that involve human labeling, such as subjective quality assessment. The target of subjective quality experiment is to obtain human opinions on the quality of visual contents affected by factors that would either enhance or degrade the visual quality with the belief that human perception is the ultimate standard. Since the human reviewing process usually comes with high financial cost and is time-consuming, the source contents allowed in subjective experiments are often extremely limited. Therefore, it is highly desirable that the selection of source content candidates are representative in terms of human viewing experience. Video quality expert group [115] has been the organization producing video content databases for the subjective experiments serving compression quality purpose. The source contents of those databases are selected based on the so called "expert experience" by experts to cover different types of content based merely on several experts' personal judgment. However, with the limited number of source contents selected with personal bias, the representativeness of the content selected remains unchecked, and the resulting uncertain impact on the subjective testing results and subsequent learning and testing experiments is unmanageable.

It can be observed that the best matches of the proposed maximum representativeness content selection methodology are those applications where abundant data are easily accessible but in need of human judgement on the data selection process. Beyond the subjective experiment for visual quality, there are many opportunities in other research areas that are in need of representative selection. One example is visual aesthetic assessment for images, which has a similar situation to image or video quality assessment. The available visual contents online for aesthetic assessment are countless, while human selection may result in selection bias, leading to less representative subjective experiment results. The maximum-representativeness selection methodology may also find applications in research area beyond the signal processing field. For example, online reviews are common nowadays. For some popular products, the number of customer reviews may be so large that are almost impossible for a single consumer to go through all of them before the consumer can obtain a general idea of the customer feedback on the product. The proposed repre-

sentative selection may provide a solution on helping consumers to find reviews that are most representative and useful in comparing the products they intend to purchase.

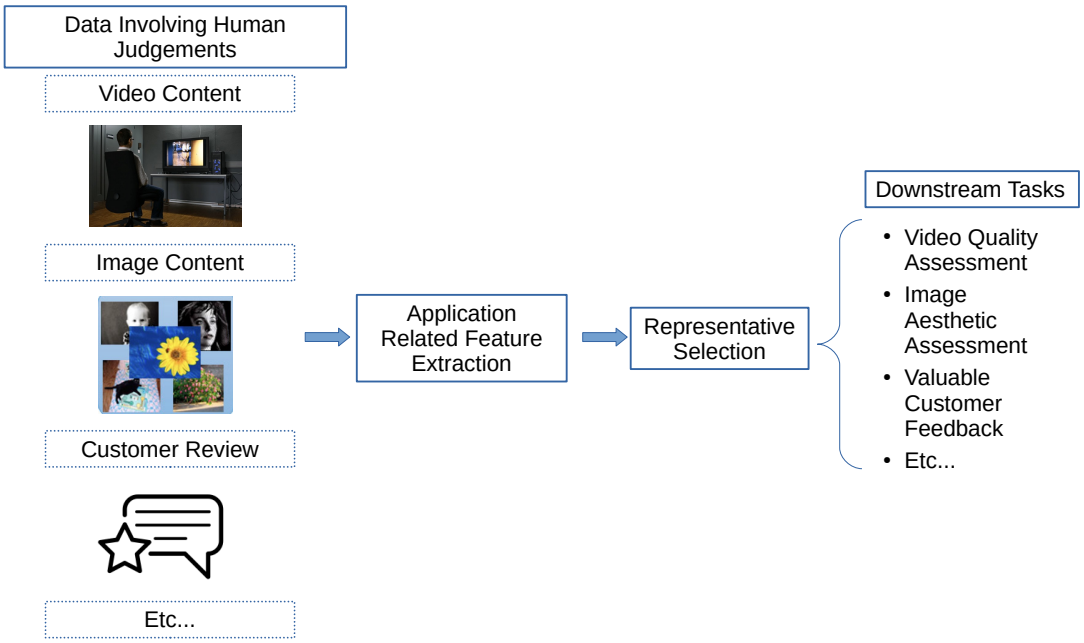


Figure 6.3: Examples of Potential Tasks for Representative Selection

In Fig.6.3, a general workflow of representative selection for the aforementioned candidate tasks is shown. Based on the nature of different applications, the application related features should be collected from the large data collection available. An example of such application related features is the RD statistic from source visual content for compression task. In the space of these features, representative selection can be conducted using the submodular optimization method described in the thesis. In the end, the selected representative data points can be utilized in the aforementioned downstream tasks. It should be noted that efforts need to be made on validating the method's effectiveness for subjective experiment. Unlike the learning based task in the thesis, the effectiveness of submodular representative selection on subjective experiments cannot be proved without human par-

ticipation. In order to prove its effectiveness, one possible solution is to utilize a much larger scale subjective experiment as the ground-truth to see if its final result correlates well with the small scale submodular based subjective experiment. In case it is difficult to accommodate the large scale subjective experiment in lab-setting, crowd-sourcing may be an alternative for the validation experiment.

In summary, the proposed methodology in the thesis provides a novel approach to measure representativeness expressed in a target optimization function for the submodular selection procedure and therefore interpret the corresponding problem as a combinatorial discrete optimization problem. The philosophy of maximum representativeness selection proposed in the thesis has broad future perspectives not only for visual quality related tasks, but also for other signal processing tasks and beyond. I hope the study in this thesis can facilitate future research on the topic, and help address the long standing problem on how to improve the representativeness in data selection that facilitates various learning-based but data-hungry signal processing applications.

References

[1] Z. Duanmu, W. Liu, Z. Li, K. Ma, and Z. Wang, "Characterizing generalized rate-distortion performance of video coding: An eigen analysis approach," *IEEE Transactions on Image Processing*, vol. 29, pp. 6180–6193, 2020.

[2] S. Athar, Z. Wang, and Z. Wang, "Deep neural networks for blind image quality assessment: Addressing the data challenge," *arXiv preprint arXiv:2109.12161*, 2021.

[3] H. Yu and S. Winkler, "Image complexity and spatial information," in *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2013, pp. 12–17.

[4] ITU-R BT.910, "Recommendation: Subjective video quality assessment methods for multimedia applications," Apr. 2008.

[5] Z. He, J. Cai, and C. W. Chen, "Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding," *IEEE Transactions on circuits and systems for video technology*, vol. 12, no. 6, pp. 511–523, 2002.

[6] S. Ma, W. Gao, and Y. Lu, "Rate-distortion analysis for h. 264/avc video coding and its application to rate control," *IEEE transactions on circuits and systems for video technology*, vol. 15, no. 12, pp. 1533–1544, 2005.

[7] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "Ssim-motivated rate-distortion optimization for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 4, pp. 516–529, 2011.

[8] A. Rehman, K. Zeng, and Z. Wang, "Display device-adapted video quality-of-experience assessment," in *Human Vision and Electronic Imaging XX*, vol. 9394, no. 939406, 2015, pp. 1–11.

[9] Z. Duanmu, W. Liu, and Z. Wang, "Modeling generalized rate-distortion functions," *arXiv preprint arXiv:1906.05178*, Jun. 2019.

[10] K. Wei, R. Iyer, and J. Bilmes, "Submodularity in data subset selection and active learning," in *International conference on machine learning*. PMLR, 2015, pp. 1954–1963.

[11] J. Bilmes, "Submodularity in machine learning and artificial intelligence," *arXiv preprint arXiv:2202.00132*, 2022.

[12] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *Int'l Conf on Learning Representations (ICLR)*, Toulon, France, April 2017.

[13] I. I. Groen, E. H. Silson, and C. I. Baker, "Contributions of low-and high-level properties to neural processing of visual scenes in the human brain," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1714, p. 20160102, 2017.

[14] R. I.-R. BT *et al.*, "Studio encoding parameters of digital television for standard 4: 3 and wide-screen 16: 9 aspect ratios," *Int. Radio Consultative Committee Int. Telecommun. Union, Switzerland, CCIR Rep*, pp. 624–4, 2011.

[15] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, and J. Häkkinen, "Cid2013: A database for evaluating no-reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 390–402, 2014.

[16] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[17] 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*, vol. 30, pp. 1–9, 2015.

[18] Y. Wang, S. Inguva, and B. Adsumilli, "Youtube ugc dataset for video compression research," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–5.

[19] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The konstanz natural video database (konvid-1k)," in *2017 Ninth international conference on quality of multimedia experience (QoMEX)*. IEEE, 2017, pp. 1–6.

[20] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612–627, 2018.

[21] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Ugc-vqa: Benchmarking blind video quality assessment for user generated content," *IEEE Transactions on Image Processing*, vol. 30, pp. 4449–4464, 2021.

[22] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Human vision and electronic imaging VIII*, vol. 5007. SPIE, 2003, pp. 87–95.

[23] W. Sun, F. Zhou, and Q. Liao, "Mdid: A multiply distorted image database for image quality assessment," *Pattern Recognition*, vol. 61, pp. 153–168, 2017.

[24] A. Katsenou, F. Zhang, M. Afonso, G. Dimitrov, and D. R. Bull, "Bvi-cc: A dataset for research on video compression and quality assessment," *Frontiers in Signal Processing*, p. 22, 2020.

[25] A. Forsythe, G. Mulhern, and M. Sawey, "Confounds in pictorial sets: The role of complexity and familiarity in basic-level picture processing," *Behavior research methods*, vol. 40, no. 1, pp. 116–129, 2008.

[26] M. M. Marin and H. Leder, "Examining complexity across domains: relating subjective and objective measures of affective environmental scenes, paintings and music," *PloS one*, vol. 8, no. 8, p. e72412, 2013.

[27] P. Machado, J. Romero, M. Nadal, A. Santos, J. Correia, and A. Carballal, "Computerized measures of visual complexity," *Acta psychologica*, vol. 160, pp. 43–57, 2015.

[28] L. Ming and P. M. Vitányi, "Kolmogorov complexity and its applications," in *Algorithms and complexity*. Elsevier, 1990, pp. 187–254.

[29] P. Machado and A. Cardoso, "Computing aesthetics," in *Brazilian symposium on artificial intelligence*. Springer, 1998, pp. 219–228.

[30] J. Romero, P. Machado, A. Carballal, and A. Santos, "Using complexity estimates in aesthetic image classification," *Journal of Mathematics and the Arts*, vol. 6, no. 2-3, pp. 125–136, 2012.

[31] D. Durmus, "Spatial frequency and the performance of image-based visual complexity metrics," *IEEE Access*, vol. 8, pp. 100 111–100 119, 2020.

[32] H. Wu, C. Mark, and K. Robert, "A study of video motion and scene complexity," *Tech. Rep. WPI-CS-TR-06-19, Worcester Polytechnic Institute*, 2006.

[33] ITU. Recommendation ITU-R H.261. (1988) H.261 : Video codec for audiovisual services at p x 384 kbit/s. [Online]. Available: https://www.itu.int/rec/T-REC-H.261-198811-S/en

[34] ITU. Recommendation ITU-R H.262. (1995) Generic coding of moving pictures and associated audio information: Video. [Online]. Available: https://www.itu.int/rec/T-REC-H.262

[35] ITU. Recommendation ITU-R H.263. (1996) Video coding for low bit rate communication. [Online]. Available: https://www.itu.int/rec/T-REC-H.263

[36] ITU. Recommendation ITU-R H.264. (2003) Advanced video coding for generic audiovisual services. [Online]. Available: https://www.itu.int/rec/T-REC-H.264

[37] ITU. Recommendation ITU-R H.266. (2020) H.266 : Versatile video coding. [Online]. Available: https://www.itu.int/rec/T-REC-H.266

[38] Z. Wang, E. Simoncelli, A. Bovik *et al.*, "Multi-scale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*, 2003, pp. 1398–1402.

[39] G. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the high efficiency video coding standard," *IEEE Transactions on circuits and systems for video technology*, pp. 1649–1668, 2012.

[40] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h. 264/avc video coding standard," *IEEE Transactions on circuits and systems for video technology*, vol. 13, no. 7, pp. 560–576, 2003.

[41] M. Budagavi, A. Fuldseth, and G. Bjøntegaard, "Hevc transform and quantization," in *High Efficiency Video Coding (HEVC)*, 2014, pp. 141–169.

[42] A. Tabatabai, T. Suzuki, P. Hanhart, P. Korshunov, T. Ebrahimi, M. Horowitz, F. Kossentini, and H. Tmar, "Compression performance analysis in hevc," in *High Efficiency Video Coding (HEVC)*, 2014, pp. 275–302.

[43] B. Li, H. Li, L. Li, and J. Zhang, "Λdomain Rate Control Algorithm for High Efficiency Video Coding," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3841–3854, 2014.

[44] Z. Li, Z. Duanmu, W. Liu, and Z. Wang, "Avc, hevc, vp9, avs2 or av1?—a comparative study of state-of-the-art video encoders on 4k videos," in *International Conference on Image Analysis and Recognition*. Springer, 2019, pp. 162–173.

[45] C. G. Bampis, Z. Li, I. Katsavounidis, T.-Y. Huang, C. Ekanadham, and A. C. Bovik, "Towards perceptually optimized adaptive video streaming-a realistic quality of experience database," *IEEE Transactions on Image Processing*, vol. 30, pp. 5182–5197, 2021.

[46] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik, "Study of temporal effects on subjective video quality of experience," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5217–5231, 2017.

[47] J. Bienik, M. Uhrina, and P. Kortis, "Impact of constant rate factor on objective video quality assessment," *Advances in Electrical and Electronic Engineering*, vol. 15, no. 4, pp. 673–682, 2017.

[48] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. (2016, Jun.) Toward a practical perceptual video quality metric. [Online]. Available: https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[49] Werner Robitza. (2017) Crf guide (constant rate factor in x264, x265 and libvpx). [Online]. Available: https://slhck.info/video/2017/02/24/crf-guide.html

[50] VideoLAN. (Jul. 2018) x264. [Online]. Available: http://git.videolan.org/git/x264

[51] MultiCoreWare Inc. (2019) x265 home page. [Online]. Available: http://x265.org/

[52] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, 2003, pp. 1398–1402.

[53] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in neural information processing systems*, vol. 31, 2018.

[54] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "Dvc: An end-to-end deep video compression framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 006–11 015.

[55] F. Jiang, W. Tao, S. Liu, J. Ren, X. Guo, and D. Zhao, "An end-to-end compression framework based on convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3007–3018, 2017.

[56] D. Marpe, H. Schwarz, and T. Wiegand, "Context-based adaptive binary arithmetic coding in the h.264/avc video compression standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 620–636, 2003.

[57] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," in *ITU-T Q. 6/SG16, 33th VCEG Meeting*, 2001.

[58] ——, "Improvements of the BD-PSNR model, VCEG-AI11," in *ITU-T Q. 6/SG16, 34th VCEG Meeting*, 2008.

[59] T. Tan, M. Mrak, V. Baroncini, and N. Ramzan, "Report on HEVC compression performance verification testing," *Joint Collab. Team Video Coding (JCT-VC)*, 2014.

[60] Alliance for Open Media. (March. 2018) The alliance for open media kickstarts video innovation era with "AV1" release. [Online]. Available: https://aomedia.org/the-alliance-for-open-media-kickstarts-video-innovation-era-with-av1-release/

[61] PKU-VCL. (2018) AVS2 technology. [Online]. Available: http://www.avs.org.cn/avs2/technology.asp

[62] 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*, vol. 26, no. 1, pp. 98–117, 2009.

[63] S.-H. Bae, J. Kim, M. Kim, S. Cho, and J. S. Choi, "Assessments of subjective video quality on HEVC-encoded 4K-UHD video for beyond-HDTV broadcasting services," *IEEE Trans. Broadcasting*, vol. 59, no. 2, pp. 209–222, 2013.

[64] S. Deshpande, "Subjective and objective visual quality evaluation of 4K video using AVC and HEVC compression," in *SID Symposium Digest of Technical Papers*, vol. 43, no. 1, 2012, pp. 481–484.

[65] P. Hanhart, M. Rerabek, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation of the upcoming HEVC video compression standard," in *Applications of Digital Image Processing XXXV*, vol. 8499, no. 84990V, 2012, pp. 1–13.

[66] M. Řeřábek and T. Ebrahimi, "Comparison of compression efficiency between HEVC/H.265 and VP9 based on subjective assessments," in *Applications Of Digital Image Processing Xxxvii*, vol. 9217, no. 92170U, 2014, pp. 1–13.

[67] Y. Zhu, L. Song, R. Xie, and W. Zhang, "SJTU 4K video subjective quality dataset for content adaptive bit rate estimation without encoding," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, 2016, pp. 1–4.

[68] M. Cheon and J.-S. Lee, "Subjective and objective quality assessment of compressed 4K UHD videos for immersive experience," *IEEE Trans. Circuits and Systems*, vol. 28, no. 7, pp. 1467–1480, 2018.

[69] Google. (Jul. 2018) libvpx. [Online]. Available: https://chromium.googlesource.com/webm/libvpx.git

[70] Alliance for Open Media. (Jun. 2018) AV1 codec source code repository. [Online]. Available: https://aomedia.googlesource.com/aom

[71] PKU-VCL. (Jan. 2018) AVS2 codec source code repository. [Online]. Available: https://github.com/pkuvcl/xavs2

[72] MultiCoreWare Inc. (Jul. 2018) x265. [Online]. Available: https://bitbucket.org/multicoreware/x265

[73] P. Fröhlich, S. Egger, R. Schatz, M. Mühlegger, K. Masuch, and B. Gardlo, "QoE in 10 seconds: Are short video clip lengths sufficient for quality of experience assessment?" in *Proc. IEEE Int. Conf. on Quality of Multimedia Experience*, 2012, pp. 242–247.

[74] Y. LIU. (Apr. 2018) AV1 beats x264 and libvpx-vp9 in practical use case . [Online]. Available: https://code.fb.com/video-engineex265ring/av1-beats-x264-and-libvpx-vp9-in-practical-use-case/

[75] ITU-R BT.500, "Recommendation: Methodology for the subjective assessment of the quality of television pictures," Jan. 2012.

[76] K. Ma, Q. Wu, Z. Wang, Z. Duanmu, H. Yong, H. Li, and L. Zhang, "Group MAD competition-A new methodology to compare objective image quality models," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 1664–1673.

[77] J. Lainema and W. Han, "Intra-picture prediction in HEVC," in *High Efficiency Video Coding (HEVC)*, 2014, pp. 91–112.

[78] Y. Chen, D. Murherjee, J. Han, A. Grange, Y. Xu, Z. Liu, S. Parker, C. Chen, H. Su, U. Joshi *et al.*, "An overview of core coding tools in the av1 video codec," in *2018 Picture Coding Symposium (PCS)*, 2018, pp. 41–45.

[79] Z. He, L. Yu, X. Zheng, S. Ma, and Y. He, "Framework of avs2-video coding," in *2013 IEEE International Conference on Image Processing*, 2013, pp. 1515–1519.

[80] P. Chen, "Video coding using extended block sizes," in *ITU-T Q. 6/SG16, 35th VCEG Meeting, USA,(Oct. 2008)*, 2008.

[81] S. Ma and C.-C. J. Kuo, "High-definition video coding with super-macroblocks," in *Visual Communications and Image Processing 2007*, vol. 6508. International Society for Optics and Photonics, 2007, p. 650816.

[82] S.-i. Sekiguchi and S. Yamagishi, "On coding efficiency with extended block size for uhdtv," *MPEG document M*, vol. 16019, 2009.

[83] S. Radicke, J.-U. Hahn, C. Grecos, and Q. Wang, "A multi-threaded full-feature hevc encoder based on wavefront parallel processing," in *2014 International Conference on Signal Processing and Multimedia Applications (SIGMAP)*. IEEE, 2014, pp. 90–98.

[84] J. Ozer. (2019) Good News: AV1 Encoding Times Drop to Near-Reasonable Levels . [Online]. Available: http://old.streamingmedia.com/Articles/ReadArticle.aspx?ArticleID=130284&PageNum=2

[85] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[86] B. Bross, J. Chen, J.-R. Ohm, G. J. Sullivan, and Y.-K. Wang, "Developments in international video coding standardization after avc, with an overview of versatile video coding (vvc)," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1463–1493, 2021.

[87] Y. Li, Z. Liu, Z. Chen, and S. Liu, "Rate control for versatile video coding," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 1176–1180.

[88] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[89] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[90] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7939–7948.

[91] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 221–231.

[92] Y. Choi, M. El-Khamy, and J. Lee, "Variable rate deep image compression with a conditional autoencoder," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3146–3154.

[93] L. Zhou, C. Cai, Y. Gao, S. Su, and J. Wu, "Variational autoencoder for low bit-rate image compression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2617–2620.

[94] I. Ahmad, V. Swaminathan, A. Aved, and S. Khalid, "An overview of rate control techniques in hevc and shvc video encoding," *Multimedia Tools and Applications*, pp. 1–32, 2021.

[95] D. Liu, Y. Li, J. Lin, H. Li, and F. Wu, "Deep learning-based video coding: A review and a case study," *ACM Computing Surveys (CSUR)*, vol. 53, no. 1, pp. 1–35, 2020.

[96] Z. Duanmu, K. Ma, and Z. Wang, "Quality-of-experience of adaptive video streaming: Exploring the space of adaptations," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1752–1760.

[97] T. Laude, Y. G. Adhisantoso, J. Voges, M. Munderloh, and J. Ostermann, "A comprehensive video codec comparison," *APSIPA Transactions on Signal and Information Processing*, vol. 8, 2019.

[98] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[99] Z. Li, A. Aaron, L. Katsavounidis, A. Moorthy, and M. Manohara. (2016) Toward a practical perceptual video quality metric. [Online]. Available: http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html

[100] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[101] F. Zhang, F. M. Moss, R. Baddeley, and D. R. Bull, "Bvi-hd: A video quality database for hevc compressed and texture synthesized content," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2620–2630, 2018.

[102] T. Li, X. Min, H. Zhao, G. Zhai, Y. Xu, and W. Zhang, "Subjective and objective quality assessment of compressed screen content videos," *IEEE Transactions on Broadcasting*, vol. 67, no. 2, pp. 438–449, 2021.

[103] P. Akyazi and T. Ebrahimi, "Comparison of compression efficiency between hevc/h.265, vp9 and av1 based on subjective quality assessments," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, 2018, pp. 1–6.

[104] E. Kreyszig, *Introductory Functional Analysis with Applications*. Wiley New York, 1978.

[105] 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.

[106] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.

[107] F. Gao, D. Tao, X. Gao, and X. Li, "Learning to rank for blind image quality assessment," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 10, pp. 2275–2290, Oct. 2015.

[108] T. George, S. Wenzhe, T. Radu, T. Lucas, B. Johannes, A. Eirikur, J. Nick, and M. Fabian, "Workshop and challenge on learned image compression (clic2020)," CVPR, 2020. [Online]. Available: http://www.compression.cc

[109] Y. Li, B. Li, D. Liu, and Z. Chen, "A convolutional neural network-based approach to rate control in hevc intra coding," in *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017, pp. 1–4.

[110] X. Lu, B. Zhou, X. Jin, and G. R. Martin, "A rate control scheme for hevc intra coding using convolution neural network (cnn)." in *DCC*, 2020, p. 382.

[111] P. Helle, H. Schwarz, T. Wiegand, and K.-R. Müller, "Reinforcement learning for video encoder control in hevc," in *2017 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2017, pp. 1–5.

[112] J.-H. Hu, W.-H. Peng, and C.-H. Chung, "Reinforcement learning for hevc/h. 265 intra-frame rate control," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2018, pp. 1–5.

[113] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual review of neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.

[114] Sandeep Kumar. (2022) How many images are on the internet in 2022? [Online]. Available: https://www.16best.net/how-many-images-are-on-the-internet/

[115] VQEG. (2022) Vqeg homepage. [Online]. Available: https://vqeg.org/vqeg-home/

[116] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 1004–1016, 2016.

[117] J. R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards—including high efficiency video coding (HEVC)," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 12, pp. 1669–1684, 2012.

[118] D. Grois, D. Marpe, A. Mulayoff, B. Itzhaky, and O. Hadar, "Performance comparison of h. 265/mpeg-hevc, vp9, and h. 264/mpeg-avc encoders," in *2013 Picture Coding Symposium (PCS)*. IEEE, 2013, pp. 394–397.

[119] Y.-F. Ou, Y. Xue, and Y. Wang, "Q-star: A perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions," *IEEE Transactions on Image Processing*, vol. 23, no. 6, pp. 2473–2486, 2014.

[120] W. Zhang, Y. Wen, Z. Chen, and A. Khisti, "Qoe-driven cache management for http adaptive bit rate streaming over wireless networks," *IEEE transactions on multimedia*, vol. 15, no. 6, pp. 1431–1445, 2013.

[121] L. Toni, R. Aparicio-Pardo, K. Pires, G. Simon, A. Blanc, and P. Frossard, "Optimal selection of adaptive streaming representations," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 11, no. 2s, pp. 1–26, 2015.

[122] Z. Wang, K. Zeng, A. Rehman, H. Yeganeh, and S. Wang, "Objective video presentation qoe predictor for smart adaptive video streaming," in *Applications of Digital Image Processing XXXVIII*, vol. 9599. SPIE, 2015, pp. 311–323.

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

[124] P. B. Mirchandani and R. L. Francis, *Discrete location theory*. Wiley, New York, 1990.

[125] J. M. Schreiber, J. A. Bilmes, and W. S. Noble, "apricot: Submodular selection for data summarization in python." *J. Mach. Learn. Res.*, vol. 21, pp. 161–1, 2020.

[126] G. K. Wallace, "The jpeg still picture compression standard," *Communications of the ACM*, vol. 34, no. 4, pp. 30–44, 1991.

[127] M. Rabbani and R. Joshi, "An overview of the jpeg 2000 still image compression standard," *Signal processing: Image communication*, vol. 17, no. 1, pp. 3–48, 2002.

[128] Alliance for Open Media. (2022) Avif homepage. [Online]. Available: https://aomediacodec.github.io/av1-avif/

[129] M. Hannuksela, E. Aksu, V. M. Vadakital, and J. Lainema, "Overview of the high efficiency image file format," *JCTVC-V0072*, 2015.

[130] Google Inc. (2022) Webp homepage. [Online]. Available: https://developers.google.com/speed/webp

[131] D. Wang, F. Speranza, A. Vincent, T. Martin, and P. Blanchfield, "Toward optimal rate control: a study of the impact of spatial resolution, frame rate, and quantization on subjective video quality and bit rate," in *Visual Communications and Image Processing 2003*, vol. 5150, 2003, pp. 198–209.

[132] T. K. Tan, R. Weerakkody, M. Mrak, N. Ramzan, V. Baroncini, J. Ohm, and G. J. Sullivan, "Video quality evaluation methodology and verification testing of hevc compression performance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 76–90, 2016.

[133] R. Yang, R. Timofte, M. Zheng, Q. Xing, M. Qiao, M. Xu, L. Jiang, H. Liu, Y. Chen, Y. Ben *et al.*, "Ntire 2022 challenge on super-resolution and quality enhancement of compressed video: Dataset, methods and results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1221–1238.

[134] J. Mukherjee, K. Praveen, and V. Madumbu, "Visual quality enhancement of images under adverse weather conditions," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 3059–3066.

[135] R. Yang, M. Xu, and Z. Wang, "Decoder-side hevc quality enhancement with scalable convolutional neural network," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 817–822.

[136] Cisco Corporation. (2018) Cisco visual networking index: Forecast and trends, 2017–2022. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html

[137] Ericsson Corporation. (2022) Ericsson mobility report. [Online]. Available: https://www.ericsson.com/49d3a0/assets/local/reports-papers/mobility-report/documents/2022/ericsson-mobility-report-june-2022.pdf

[138] R. M. Nasiri, J. Wang, A. Rehman, S. Wang, and Z. Wang, "Perceptual quality assessment of high frame rate video," in *2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP)*, 2015, pp. 1–6.

[139] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[140] ISO/IEC 11172-2:1993. (1993) Coding of moving pictures and associated audio information for digital storage media at up to about 1.5 Mbit/s – Part 2: Video. (mpeg-1). [Online]. Available: https://www.iso.org/standard/22411.html

[141] ISO/IEC 14496-2:1999. (1999) Coding of audio-visual objects – Part 2: Video. (mpeg-4 visual). [Online]. Available: https://www.iso.org/standard/25034.html

[142] H. Schwarz, T. Schierl, and D. Marpe, "Block structures and parallelism features in HEVC," in *High Efficiency Video Coding (HEVC)*, 2014, pp. 49–90.

[143] L. Shen, Z. Zhang, and P. An, "Fast CU size decision and mode decision algorithm for HEVC intra coding," *IEEE Transactions on Consumer Electronics*, vol. 59, no. 1, pp. 207–213, 2013.

[144] S. Cho and M. Kim, "Fast CU splitting and pruning for suboptimal CU partitioning in HEVC intra coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 9, pp. 1555–1564, 2013.

[145] G. Chen, Z. Liu, T. Ikenaga, and D. Wang, "Fast HEVC intra mode decision using matching edge detector and kernel density estimation alike histogram generation," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*, 2013, pp. 53–56.

[146] T. L. Da S., L. V. Agostini, and L. A. da Silva Cruz, "Fast HEVC intra prediction mode decision based on EDGE direction information," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 1214–1218.

[147] W. Jiang, H. Ma, and Y. Chen, "Gradient based fast mode decision algorithm for intra prediction in HEVC," in *2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet)*, 2012, pp. 1836–1840.

[148] L. Zhao, L. Zhang, S. Ma, and D. Zhao, "Fast mode decision algorithm for intra prediction in HEVC," in *2011 Visual Communications and Image Processing (VCIP)*, 2011, pp. 1–4.

[149] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, pp. 114–117, 1965.

[150] L. Eeckhout, "Is Moore's Law Slowing Down? What's Next?" *IEEE Micro*, no. 4, pp. 4–5, 2017.

[151] J. Y. Chen and J. E. Thropp, "Review of low frame rate effects on human performance," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 37, no. 6, pp. 1063–1076, 2007.

[152] T. Mallikarachchi, D. S. Talagala, H. K. Arachchi, and A. Fernando, "Content-adaptive feature-based CU size prediction for fast low-delay video encoding in HEVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 693–705, 2016.

[153] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of Go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, p. 484, 2016.

[154] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of Go without human knowledge," *Nature*, vol. 550, no. 7676, p. 354, 2017.

[155] S. Ma, W. Gao, and Y. Lu, "Rate-Distortion Analysis for H.264/AVC Video Coding and its Application to Rate Control," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 12, pp. 1533–1543, 2005.

[156] Y. Liu, Z. G. Li, and Y. C. Soh, "A novel rate control scheme for low delay video communication of H.264/AVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 1, pp. 68–78, 2007.

[157] S. Hu, H. Wang, and S. Kwong, "Adaptive quantization-parameter clip scheme for smooth quality in H.264/AVC," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1911–1919, 2012.

[158] Z. He and S. K. Mitra, "Optimum bit allocation and accurate rate control for video coding via ρ-domain source modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 10, pp. 840–849, 2002.

[159] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 23–50, 1998.

[160] Z. He, Y. K. Kim, and S. K. Mitra, "Low-delay rate control for dct video coding via/spl rho/-domain source modeling," *IEEE transactions on Circuits and Systems for Video Technology*, vol. 11, no. 8, pp. 928–940, 2001.

[161] M. Liu, Y. Guo, H. Li, and C. W. Chen, "Low-complexity rate control based on ρ-domain model for scalable video coding," in *2010 IEEE International Conference on Image Processing*. IEEE, 2010, pp. 1277–1280.

[162] C. S. Lim, S. M. T. Naing, V. Wahadaniah, and X. Jing, "Reference lists for b pictures under low delay constraints," *document JCTVC-D093, ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC), Daegu, Korea*, 2011.

[163] D.-K. Kwon, M.-Y. Shen, and C.-C. J. Kuo, "Rate control for h. 264 video with enhanced rate and distortion models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 5, pp. 517–529, 2007.

[164] Y. Liu, Z. G. Li, and Y. C. Soh, "A novel rate control scheme for low delay video communication of h. 264/avc standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 1, pp. 68–78, 2006.

[165] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, p. 529, 2015.

[166] F.-f. Li, J. Johnson, and S. Yeung. (2017) CS231n Reinforcement Learning. [Online]. Available: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture14.pdf

[167] F. Hayes-Roth, D. A. Waterman, and D. B. Lenat, *Building expert systems*. Addison-Wesley Longman Publishing Co., Inc., 1983.

[168] American Go Association. (2019) Top ten reasons to play go. [Online]. Available: https://www.usgo.org/top-ten-reasons-play-go

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

[170] Z. Li, T. Vigier, and P. L. Callet. (2018, Mar.) A vmaf model for 4k. [Online]. Available: ftp://vqeg.its.bldrdoc.gov/Documents/VQEG_Madrid_Mar18/Meeting_Files/VQEG_SAM_2018_025_VMAF_4K.pdf

[171] VQEG. (2000, Apr.) Final report from the video quality experts group on the validation of objective models of video quality assessment. [Online]. Available: http://www.vqeg.org/

[172] ITU-R BT.2020, "Recommendation: Parameter values for ultra-high definition television systems for production and international programme exchange," Apr. 2015.

[173] P. Massimino. (Jul. 2017) AOM - AV1, How does it work? [Online]. Available: https://parisvideotech.com/wp-content/uploads/2017/07/AOM-AV1-Video-Tech-meet-up.pdf

[174] P. Akyazi and T. Ebrahimi, "Comparison of Compression Efficiency between HEVC/H. 265, VP9 and AV1 based on Subjective Quality Assessments," in *Proc. IEEE Int. Conf. on Quality of Multimedia Experience*, 2018, pp. 1–6.

[175] J. Li, B. Li, J. Xu, and R. Xiong, "Intra prediction using fully connected network for video coding," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 1–5.

[176] Y. Hu, W. Yang, M. Li, and J. Liu, "Progressive spatial recurrent neural network for intra prediction," *IEEE Transactions on Multimedia*, 2019.

[177] Y. Wang, X. Fan, C. Jia, D. Zhao, and W. Gao, "Neural network based inter prediction for hevc," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[178] P. Helle, H. Schwarz, T. Wiegand, and K. R. Müller, "Reinforcement learning for video encoder control in HEVC," in *International Conference on Systems, Signals, and Image Processing*, 2017, pp. 1–5.

[179] L.-C. Chen, J.-H. Hu, and W.-H. Peng, "Reinforcement Learning for HEVC/H. 265 Frame-level Bit Allocation," in *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, 2018, pp. 1–5.

[180] N. Li, Y. Zhang, L. Zhu, W. Luo, and S. Kwong, "Reinforcement learning based coding unit early termination algorithm for high efficiency video coding," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 276–286, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320319300677

[181] C. Chung, W. Peng, and J. Hu, "HEVC/H.265 coding unit split decision using deep reinforcement learning," in *2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, 2017, pp. 570–575.

[182] J.-H. Hu, W.-H. Peng, and C.-H. Chung, "Reinforcement learning for hevc/h. 265 intra-frame rate control," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018, pp. 1–5.

[183] R. B. Segal, "On the scalability of parallel uct," in *International Conference on Computers and Games*, 2010, pp. 36–47.

[184] A. Saxena and F. C. Fernandes, "Dct/dst-based transform coding for intra prediction in image/video coding," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3974–3981, 2013.

[185] IEEE Communications Society. (2016) P1918.1-Tactile Internet: Application Scenarios, Definitions and Terminology, Architecture, Functions, and Technical Assumptions. [Online]. Available: https://standards.ieee.org/project/1918_1.html

[186] R. Hassen, B. Gulecyuz, and E. Steinbach, "PVC-SLP: Perceptual Vibrotactile-Signal Compression based-on Sparse Linear Prediction," *IEEE Transaction on Multimedia (Early Access)*, 2020.

[187] R. L. Klatzky and S. J. Lederman, "Touch," in *Handbook of psychology, Volume 4: Experimental psychology*. John Wiley &Sons, Inc, 2003, ch. 6, pp. 147–176.

[188] S. Okamoto, H. Nagano, and Y. Yamada, "Psychophysical dimensions of tactile perception of textures," *IEEE Transactions on Haptics*, vol. 6, no. 1, pp. 81–93, 2013.

[189] L. R. Manfredi, H. P. Saal, K. J. Brown, M. C. Zielinski, J. F. Dammann III, V. S. Polashock, and S. J. Bensmaïa, "Natural scenes in tactile texture," *Journal of Neurophysiology*, vol. 111, no. 9, pp. 1792–1802, 2014.

[190] A. I. Weber, H. P. Saal, J. D. Lieber, J. W. Cheng, L. R. Manfredi, J. F. Dammann, and S. J. Bensmaia, "Spatial and temporal codes mediate the tactile perception of natural textures," *Proceedings of the National Academy of Sciences*, p. 201305509, 2013.

[191] S. J. Bensmaïa and M. Hollins, "The vibrations of texture," *Somatosensory & Motor Research*, vol. 20, no. 1, 2003.

[192] S. Bensmaïa and M. Hollins, "Pacinian representations of fine surface texture," *Perception & Psychophysics*, vol. 67, no. 5, pp. 842–854, 2005.

[193] O. Holland, E. Steinbach, R. V. Prasad, Q. Liu, Z. Dawy, A. Aijaz, N. Pappas, K. Chandra, V. S. Rao, S. Oteafy, M. Eid, M. Luden, A. Bhardwaj, X. Liu, J. Sachs, and J. Araújo, "The IEEE 1918.1 "Tactile Internet" Standards Working Group and Its Standards," *Proceedings of the IEEE*, pp. 1–24, 2019.

[194] E. Steinbach, M. Strese, M. Eid, X. Liu, A. Bhardwaj, Q. Liu, M. Al-Ja'afreh, T. Mahmoodi, R. Hassen, A. E. Saddik, and O. Holland, "Haptic codecs for the tactile internet," *Proceedings of the IEEE*, pp. 1–24, 2018.

[195] S. Choi and K. J. Kuchenbecker, "Vibrotactile display: Perception, technology, and applications," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2093–2104, 2013.

[196] J. M. Romano and K. J. Kuchenbecker, "Creating realistic virtual textures from contact acceleration data," *IEEE Transactions on Haptics*, vol. 5, no. 2, pp. 109–119, April 2012.

[197] K. J. Kuchenbecker, J. Romano, and W. McMahan, "Haptography: Capturing and recreating the rich feel of real surfaces," in *Robotics Research*. Springer, 2011, pp. 245–260.

[198] R. Hassen and E. Steinbach, "Vibrotactile signal compression based on sparse linear prediction and human tactile sensitivity function," in *IEEE World Haptics Conference (WHC)*, 2019.

[199] A. Noll, B. Gulecyuz, A. Hofmann, and E. Steinbach, "A rate-scalable perceptual wavelet-based vibrotactile codec," in *IEEE Haptics Symposium*, March 2020.

[200] R. Chaudhari, B. Çizmeci, K. J. Kuchenbecker, S. Choi, and E. Steinbach, "Low bitrate source-filter model based compression of vibrotactile texture signals in haptic teleoperation," in *Proceedings of the 20th ACM International Conference on Multimedia*, 2012, pp. 409–418.

[201] R. Chaudhari, C. Schuwerk, M. Danaei, and E. Steinbach, "Perceptual and bitrate-scalable coding of haptic surface texture signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 3, pp. 462–473, April 2015.

[202] S. Okamoto and Y. Yamada, "Lossy data compression of vibrotactile material-like textures," *IEEE Transactions on Haptics*, vol. 6, no. 1, pp. 69–80, 2013.

[203] ——, "Perceptual properties of vibrotactile material texture: Effects of amplitude changes and stimuli beneath detection thresholds," in *IEEE International Symposium on System Integration*, 2010, pp. 384–389.

[204] R. Hassen and E. Steinbach, "Subjective Evaluation of the Spectral Temporal SIMilarity (ST-SIM) Measure for Vibrotactile Quality Assessment," *IEEE Transactions on Haptics*, vol. 13, no. 1, pp. 25–31, 2020.

[205] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[206] J. Hale. (2019) More than 500 hours of content are now being uploaded to youtube every minute. [Online]. Available: https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/

[207] J. Perkiö and A. Hyvärinen, "Modelling image complexity by independent component analysis, with application to content-based image retrieval," in *International Conference on Artificial Neural Networks*. Springer, 2009, pp. 704–714.

[208] T. Guha and R. K. Ward, "Image similarity using sparse representation and compression distance," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 980–987, 2014.

[209] R. Cilibrasi and P. M. Vitányi, "Clustering by compression," *IEEE Transactions on Information theory*, vol. 51, no. 4, pp. 1523–1545, 2005.

[210] Z. Islam, M. Abdel-Aty, Q. Cai, and J. Yuan, "Crash data augmentation using variational autoencoder," *Accident Analysis & Prevention*, vol. 151, p. 105950, 2021.

[211] H. Nishizaki, "Data augmentation and feature extraction using variational autoencoder for acoustic modeling," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 1222–1227.

[212] C. Chadebec, E. Thibeau-Sutre, N. Burgos, and S. Allassonnière, "Data augmentation in high dimensional low sample size setting using a geometry-based variational autoencoder," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.