Video Quality Assessment: Exploring the Impact of Frame Rate

by

Rasoul Mohammadi Nasiri

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

Waterloo, Ontario, Canada, 2019

Examining Committee Membership

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

External Examiner:	Xiao-Ping Zhang Professor, Dept. of Elec. & Comp. Engineering, Ryerson University
Supervisor(s):	Zhou Wang Professor, Dept. Elec. & Comp. Engineering, University of Waterloo
Internal Member:	George Freeman Professor, Dept. of Elec. & Comp. Engineering, University of Waterloo
Internal Member:	Fakhreddine Karray Professor, Dept. of Elec. & Comp. Engineering, University of Waterloo
Internal-External Member:	John Zelek Professor, Dept. of Sys. Des. Engineering, University of Waterloo

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

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

Abstract

Technology advancements in the past decades has led to an immense increase in data traffic over various networks. Videos constitute a major percentage of this traffic and their share is projected to increase at an accelerating speed in the coming years. Service providers aim to deliver videos that have high quality while at the same time keeping the data rate as low as possible. Effective and efficient objective Video Quality Assessment (VQA) algorithms are essential in order to provide real time estimate of video quality so that the best compromise between data rate and quality can be achieved. Data rate of video transmission can be altered by controlling different parameters of the video, among which frame rate is one of the most important parameters. So far, only limited works have been done to study the impact of frame rate variations on video quality.

The purpose of this work is to investigate the impact of varying frame rate on the quality of videos and to develop novel VQA models that integrate frame rate variations into the task of quality assessment. In order to achieve this goal, we first construct two new video databases that contain videos of diverse content, and spatial and temporal resolutions. We carry out subjective studies on these databases to obtain human opinions on video quality. The subjective study allows us to evaluate the performance of well known objective VQA algorithms on cross-frame rate videos. The results reveal that there is considerable disparity between the subjective scores and the predictions from state-of-the-art objective models that do not take frame rate into consideration.

We explore statistical models for video quality analysis. In particular, we conduct cross-frame local phase statistical analysis, which provides new insights on video motion smoothness as an important factor that affects video quality across different frame rates. Our evaluations of the proposed motion smoothness metric using the subject-rated databases show that this novel measure provides a new means to capture the impact of frame rate on video quality, and demonstrates strong promise at improving the performance of objective video quality assessment models.

We also propose the notions of perceptual temporal aliasing factor and perceptual spatiotemporal aliasing factor to incorporate the characteristics of human visual contrast sensitivity variations into the framework of spatial and temporal aliasing analysis. We incorporate the proposed aliasing factors into the VQA process to predict the quality of video under frame rate change, resolution change, and lossy compression. Our performance evaluation using the subjective database shows that the proposed perceptual aliasing factors are strong quality predictors across-frame rate, resolution, and data rate, significantly outperforming baseline VQA methods without aliasing modeling.

Acknowledgements

First, thanks God for everything in my life. Special thanks to my knowledgable, genius, and encouraging supervisor, Professor Zhou Wang who dedicated too many hours during the last five years on my project. This work would not have been possible without his support. He taught me all the details of doing perfect research. He was completely understanding supervisor made this way easy for me. I always learned new things anytime I talked to him.

Thank you to my thesis committee, Professor George Freeman, Professor Fakhreddine Karray, Professor John Zelek, and Professor Xiao-Ping Zhang. I express my sincere gratitude for all your help and valuable comments made this thesis more strong.

While I had the chance of enjoying the innovative environment at the University of Waterloo, my special opportunity was to work with amazing groupmates at the Image and Vision Computing Lab. I would like to thanks them all including Shahrukh Athar, Jiheng Wang, Abdul Rehman, Zhengfang Duanmu, Shiqi Wang, Wentao Liu, Hojatollah Yeganeh, Kede Ma, Zhuoran Li, Kai Zeng, Thilan Costa, Zhongling Wang.

Last but not least, I lovely dedicate this thesis to my lovely wife, Mahzar Eisapour, for her patience, help, and encouragement in this path. She was always my best support even in her busiest days during her study at the University of Waterloo. Through her love and belief in me, I could complete this journey. No word can express my love to you.

Dedication

To my wife for her endless companionship, support, encouragement, and sacrifice,

To my parents for and all their unconditional support.

Contents

List of Tables

List of Figures

List of Abbreviations

CSF Constrast Sensitivity Function. 99, 101

CV Circular Variance. 65–67

DCT Discrete Cosine Transform. 56

DFT Discrete Fourier Transform. 89

fps Frame Per Second. 2, 3, 14, 15, 18, 19, 25, 26, 40, 43, 44, 61, 129

FR-VQA Full Reference Video Quality Assessment. 12, 13

HD High Definition. 40

HVS Human Visual System. 2, 6–8

IQA Image Quality Assessment. 5–9, 13, 22, 55, 56

IW-SSIM Information Weighted SSIM. 8

kbps Kilo bits per second. 17

KRCC Kendall's Rank Correlation Coefficient. 21, 22

MAE Mean Absolute Error. 7, 20, 21

MOS Mean Opinion Score. xviii, 7, 17, 20–22, 29, 30, 55, 56, 103, 113–115

MS-SSIM Multi-Scale SSIM. xv, 8, 13, 55, 56, 60, 110, 111, 113, 114, 128, 130

MSE Mean Squared Error. 7, 8, 13

NR-VQA No Reference Video Quality Assessment. 12, 13

NSS Natural Scene Statistics. 62, 86

PLCC Pearson Linear Correlation Coefficient. 21, 22, 114

PSNR Peak Signal to Noise Ratio. 7, 8, 13, 16, 17

QoE Quality of Experience. 18, 78

QP Quantizatin Parameter. 17, 19, 26, 44

RMSE Root Mean Squared Error. 7, 21, 114

RR-VQA Reduced Reference Video Quality Assessment. 12, 13

SD Standard Definition. 40

SI Spatial Information. 17

SRCC Spearman's Rank Correlation Coefficient. 22, 114

SSIM Structural Similarity. xv, 8, 13, 14, 16, 55, 56, 58, 114

SVR Support Vector Regression. 110, 111, 113, 114

TI Temporal Information. 17

VQA Video Quality Assessment. xii, 2, 5, 6, 9, 11–14, 17, 19, 20, 22, 24, 29, 30, 41, 55–57, 61, 63, 86, 107, 109, 110, 114, 128–131

VQM Video Quality Metric. xv, 55, 56, 59, 114

Chapter 1

Introduction

1.1 Motivation

Today videos compose a majority of data traffic over various networks. It is predicted that more than 80 percent of the traffic over the Internet would be composed of videos by 2021 [5, 6]. Such a gigantic amount of data is the result of the fast advancement of video capturing, delivery and display technologies. In particular, a great number of video capturing devices including mobile phones and personal SLR cameras have been spread to common consumers, and capturing images and videos has become part of their everyday life. The increased access to online videos is another important factor that drives the significant increase of video production and distribution.

In the past decade, the industry has made tremendous effort to build capable yet inexpensive devices for capturing and displaying videos. Each year many new models of these devices are released. Meanwhile, cable, satellite, IPTV and Internet Over-the-Top (OTT) video service providers have been striving to offer a better quality service while keeping the amount of traffic on the network as low as possible. The traffic of network may be reduced by increasing the strength of video compression, decreasing the spatial resolution, or decreasing the temporal resolution or frame rate. This allows the service providers to support more customers with the same network resources. Such data rate reduction has to be performed without loss of the perceptual quality of the videos being delivered to the end consumers. However, currently effective and efficient quality control mechanisms are largely lacking in the industry. In particular, trusted VQA methods that can be used to evaluate, compare, monitor, control and optimize the video acquisition processes, video delivery services, and video display systems in a perceptually meaningful manner are lacking in the ecosystem. This has been one of the major driving forces that have led to a remarkable growth of VQA research in recent years [7, 8, 9, 10].

VQA evaluation may be done in two ways: subjective assessment by humans or objective evaluation by computational models. Subjective quality assessment may be considered more reliable because humans are the ultimate consumers of most video services, but it also has significant drawbacks. Specifically, it is slow, expensive, and cannot be embedded into video processing systems for design and optimization purposes. Objective VQA models do not suffer from these drawbacks and have become a desired method of choice in most practical applications.

In this work we focus on the impact of frame rate on video quality. A digital video is typically represented as a sequence of 2D image frames that are discrete in time. Frame rate refers to the frequency at which the frames are displayed, and is quantified in the unit of Frame Per Second (fps). Frame rate is an important aspect of video quality, but surprisingly has rarely been deeply investigated in the literature.

Theoretically the Human Visual System (HVS) can process up to 1000 images per second but for untrained eyes, the difference is not noticeable after about 150 fps [11].

In image recognition tasks, people have been found to recognize a distinct image in an unbroken series of different images, each of which lasts as little as 13 milliseconds that leads to about 77 fps. In the industry of film making and cinema, early silent films in the first decades of the 20th century, had used frame rates between 16 and 24 fps. This rate was not fixed as the equipment were tuned manually. To avoid flickering and other effects, silent films were often intended to be shown at higher frame rates than the capturing rates. Nevertheless, some jerky motions still remained. The higher frame rates between 20 to 26 fps were used in the late 1920s. With the introduction of sound film, 24 fps became the standard to avoid changes in audio frequency and to keep audio and video synchronized. Three main frame rate standards are used in the TV and digital cinema business: 24p, 25p, and 30p. They all come from the initial standard of 24 fps with some considerations of the broadcasting technologies of PAL, SECAM, and NTSC standards. Multiple projection and interlace display techniques are applied in 24 fps video display to increase display rate and to avoid motion artifacts such as flickering and motion blur.

With the advancement of technology in film making and display devices in the last two decades, frame rates higher than the traditional 24 fps such as 48, 50, 60, and recently 120 fps have been used in cinema, television, and computer display industry. These frame rates are typically used in progressive format to avoid motion blur. Despite of such increasing use of high frame rate video, its impact on perceptual quality has rarely been deeply investigated in the literature. At this point, it still remains to be fully understood the impact of increasing frame rates more than 24 fps on human visual perception. Such understanding may play an important role in finding the best acquisition, coding and display conditions in particular application environment.

1.2 Objectives

The objectives of this research are to investigate the impact of frame rate on perceptual video quality and develop objective models that can automatically quantify and compare the quality of videos at different frame rates, so as to apply such models in the practice of video coding and display.

Specifically, considering a pristine uncompressed video that passes through a sequence of encoding steps including frame rate change to create a compressed version, the objective is to investigate the perceptual quality of the compressed video in comparison to the pristine version as a function of frame rate, and to develop objective models that automatically quantify the overall quality of the compressed video.

1.3 Contributions

The contributions of this work are summarized as follows.

- Constructed two new databases for cross-frame rate VQA and conducted two subjective studies to create a VQA benchmark.
- Proposed and evaluated a temporal motion smoothness (TMS) factor and showed its power in predicting video quality degradation due to frame rate reduction.
- Proposed and evaluated perceptual temporal and spatiotemporal aliasing factors
- Proposed a series of new VQA models based on the proposed perceptual aliasing factors to predict cross-frame rate video quality.

1.4 Organization of the Thesis

This thesis is organized as follows. In Chapter 2, We provide an overview of the field of Image Quality Assessment (IQA) and VQA, and introduce the well-known models for IQA and VQA. We then focus on the discussions about the relationship between frame rate and other video and visual characteristics such as motion perception, video quality, VQA models, and rate control in video coding. Standard evaluation criteria of objective VQA models are also discussed.

In Chapter 3, we introduce our work to create two databases and to perform subjective studies on the databases. We report the key observations from the results related to the impact of frame rate to perceptual video quality. We also investigate the performance of well-known VQA methods in predicting the video quality scores obtained from subjective studies on our video databases.

The impact of frame rate changes by considering natural scene statistics is analyzed in Chapter 4. In particular, the smoothness of motion in natural scenes is analyzed based on cross-frame rate phase correlation of complex wavelet transform. We measure the smoothness of motion in video and explore its relationship with perceptual video quality in cross frame rate videos.

In Chapter 5, the concepts of perceptual temporal aliasing factor and perceptual spatio-temporal aliasing factor are proposed by considering human visual contrast sensitivity models. The relationship between these perceptual aliasing factors and video quality degradation is also investigated. The perceptual aliasing factors are combined with frame rate blind VQA models to create comprehensive VQA models that significantly outperform baseline VQA methods that do not incorporate aliasing models.

The last chapter summarizes the current work and discusses potential future directions.

Chapter 2

Literature Review

In this chapter we first review the field of image quality assessment IQA and video quality assessment VQA, with special attention to the relationship between frame rate and video quality. We then introduce the standard criteria used to evaluate the performance of IQA and VQA models.

2.1 Image Quality Assessment

Various image processing algorithms and systems produce output images, and the performance of such algorithms and systems is evaluated by the quality of the output images that are typically consumed by human eyes. Therefore, IQA is of critical importance in evaluating these algorithms and systems.

Challenges of IQA in general have been reviewed in [12] and these include: complexity of the HVS, variety of distortions, influence of distortion on image appearance, impact of multiple distortions, geometrical distortion of image, enhanced image quality, and memory requirement. In addition, some other challenges are color handling, 3D distortions, and varying viewing conditions.

Since humans are the ultimate receivers of images, the most reliable method in IQA would be subjective assessment of image quality by human observers. One of the difficulties of subjective testing is the variations between human subjects [12]. To address this issue, the quality score of an image is often labeled by calculating the Mean Opinion Score MOS of all subjects [13]. Also, the subjective tests need to be carefully designed to collect reliable data about human opinions [14]. Efforts have been made by standard bodies to make recommendations to perform subjective test [15]. However, subjective testing is time consuming and cannot be directly embedded into IQA algorithms for optimization purposes.

Because of the fundamental limitations of subjective quality assessment, designing objective methods that can automatically estimate the quality of images is highly desirable. Traditional methods for quality assessment of signals such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Peak Signal to Noise Ratio (PSNR), and Root Mean Squared Error (RMSE) can be applied on images as well [16, 17]. However, they have been shown to poorly correlate with perceptual image quality [16]. The limitations of using MSE and PSNR to perform IQA have been presented in [16]. The author of [16] discusses why there is still very significant interest in MSE and why we need more sophisticated methods for image quality assessment. Detailed discussions about the validity of PSNR for assessing image quality are given in [17].There are also efforts to make MSE more compatible to visual perception [18].

Considering the poor performance of traditional signal quality assessment metrics in IQA, better objective IQA models should be designed to predict the quality assessment behaviors of the HVS. However, it is difficult to construct IQA algorithms that fully mimic the various functional blocks of the HVS due to the non-linearity [19, 20], complexity, and our limited knowledge about the HVS [21]. Algorithms developed under this bottom-up framework of HVS simulation have achieved only limited success and their performance is often shown to be at a similar level of MSE or PSNR but with significantly increased computational cost.

An increasingly popular IQA design principle is to consider the HVS as a black box and then to develop IQA algorithms that try to simulate the overall functionality of the HVS. The SSIM index is one of the earliest and best-known metrics in quality analysis of images follows this top-down principle [22]. This metric uses a combination of luminance, contrast, and structural distortion to arrive at an objective quality score, and meanwhile produces a quality map.

Due to the success of SSIM, different variations of SSIM have been proposed as well. One of the most well-known variations of SSIM is the MS-SSIM method proposed in [23]. In this method, a hierarchy of down-sampled images constructs a pyramid of different scales and quality degradation is calculated from the higher to the lower levels of the pyramid based on the visibility and importance of quality degradation at different levels. Information Weighted SSIM (IW-SSIM) is another well-known variation of SSIM that weights different regions in an image based on the amount of information that it contains [24]. The information content of a block is determined by applying information theoretic concepts.

With the remarkable success of deep neural networks (DNN) in many applications of computer vision [25, 26], it has been the focus of many recent studies in image quality assessment as well [27, 28, 29, 30]. While traditional IQA models using hand-crafted or learned features, DNN based IQA models may perform end-to-end quality prediction using image pixels as the input data [31, 32, 33].

2.2 Video Quality Assessment

Video quality assessment is much more difficult in comparison with IQA because it has an additional dimension of time which adds novel distortion types along the temporal direction. The visual experience in the perception of motion is another critical aspect that creates new challenges on VQA as opposed to IQA.

2.2.1 Perceptual Artifacts in Compressed Video

A distortion artifact in video is any change in the appearance of the original (high quality) video which alters the perceptual quality of video. Some of these changes are more visible and annoying than others.

Distortions can occur at different stages of creating, processing, transmitting, and displaying a video. In real-world video delivery systems, one of the most common causes of distortions is lossy video compression. Perceptual distortion artifacts that usually occur in compressed videos are summarized in [1], where distortions have been divided into two categories: spatial artifacts and temporal artifacts. Spatial artifacts are those artifacts that change the appearance of individual frames, while temporal artifacts occur during video playback. More detailed descriptions of different perceptual artifacts created by video compression are summarized in Fig. 2.1.

In spatial artifacts, blurring is usually caused by quantization of high frequency components. Blocking is usually caused by the application of block-based transforms and quantization. Another spatial artifact is ringing which occurs when the coefficients of frequency or wavelet transforms are quantized.

In temporal artifacts, flickering occurs due to frequent and repetitive changes in brightness or color [34]. Jerkiness is caused by high speed motion faster than the sampling

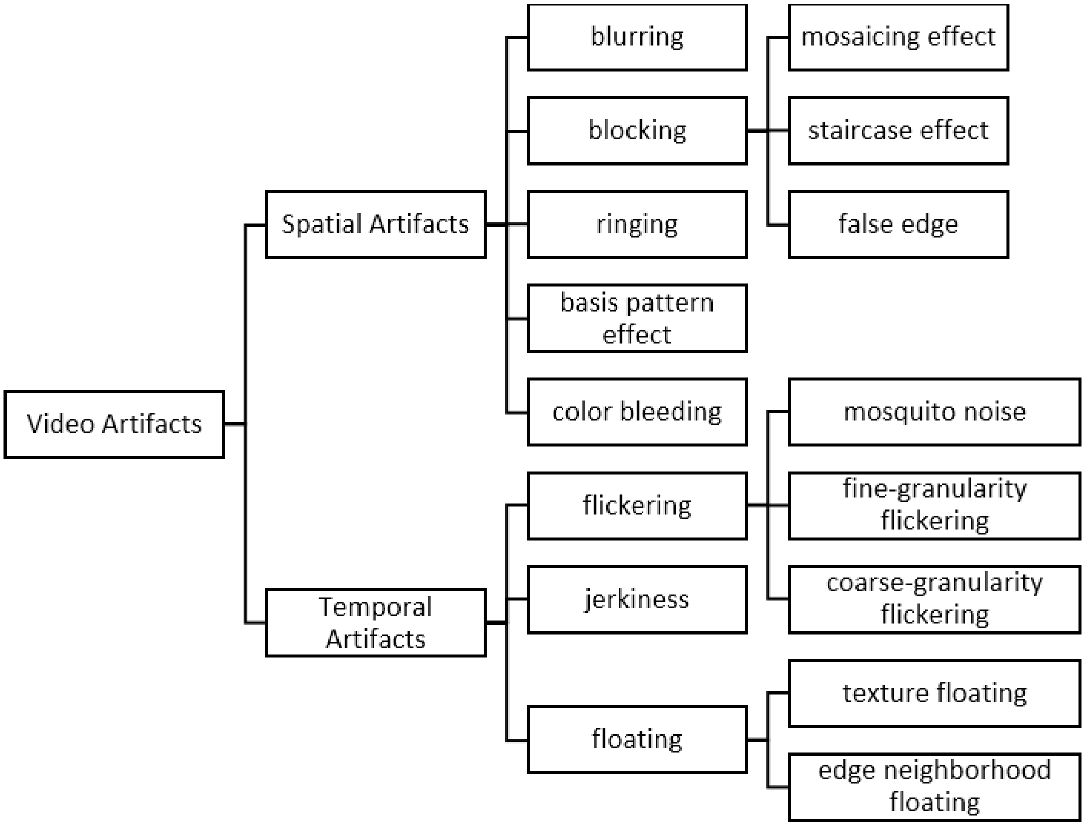


Figure 2.1: Perceptual artifacts of compressed videos [1].

frequency in time. Floating appears as a region in a frame moving in the wrong direction or at a wrong speed. This may be caused by the Skip mode in certain video compression standards, where the motion prediction residues are not coded due to limited bit budget.

2.2.2 Subjective vs Objective VQA

Traditionally the quality of videos have been evaluated by subjective tests, in which human subjects are asked to score the test video based on their perceptual quality judgments. Subjective tests are usually performed in laboratory environment. Several standards and recommendations [15, 35, 36] on how subjective tests should be conducted have been widely used in practice, where the conditions of the test environment, scoring methodology, and display conditions are defined explicitly. In particular, ITU-T P.910 [15] is a recommendation for subjective video quality assessment for general multimedia application, ITU-R BT.500 [36] is for television pictures quality tests, and ITU-T BT.710 [37] is designed for HD videos.

Depending on the presence of the pristine (undistorted) version of the test video, subjective test may be conducted using Single Stimulus (SS) or Double Stimulus (DS) approaches [15]. In SS methods the subjects view a test video and score based on perceptual quality judgments. In DS methods, the subjects view a pristine reference video first and then the test video and score the quality of the test video comparing with its pristine reference.

In a subjective test, the number of sequences should be selected to cover diverse content and to avoid a viewing session being boring for the subject. It is recommended to have at least four video sequences in each session [15].

In the subjective test session test video sequences are displayed to user one by one and after each sequence a gray screen is displayed while the subject is asked to score the perceptual quality of video on paper, or on a graphical interface, or using other scoring devices. The score is usually a number between 1-5, 0-10 or 0-100, where 5, 10 or 100 is the score for the best perceptual quality in human opinion. The categorical scoring can also be used by using 5 Likert scale questions about quality with the labels of "poor quality", "bad quality", "moderate quality", "good quality", and "excellent quality".

Subjective VQA has many drawbacks. They are costly and time-consuming. They are sensitive to test environment and conditions. They are also difficult to be used for real-time quality monitoring and performance optimization applications.

Table 2.1: Summary of pros. and cons. of Subjective vs. Objective VQA.

	Pros.	Cons.
Subjective VQA	Reliable	Expensive, Time consuming, Difficult-to-use
Objective VQA	Automatic, Low cost, Easy-to-use	Less Reliable

In contrast to subjective VQA, objective quality assessment of videos has the potentials to be performed in real time with low cost. The goal is to build objective models that make video quality predictions that correlate well with human perception. The demand for objective VQA is increasing in recent years due to the dramatical growth of video content being distributed in various communication networks.

2.2.3 Existing Objective VQA Methods

Humans may assess the quality of a test video by basing their judgements on perceptual expectations or by comparing the video with a reference video that is assumed to have pristine quality. Similarly, depending on the availability of the reference video, objective VQA models may be classified into three categories:

- Full Reference Video Quality Assessment (FR-VQA) methods
- Reduced Reference Video Quality Assessment (RR-VQA) methods
- No Reference Video Quality Assessment (NR-VQA) methods

In FR-VQA methods the original high quality video is available and is used as the reference to evaluate the quality of a distorted video (test video). In this case the quality evaluation may be performed at a precision up to pixel level in which the pixel values and/or features related to the appearance of frames are compared. The result of such local comparisons, in regions within a frame or across frames are aggregated to estimate the overall quality of the test video.

In RR-VQA methods, the original video is not fully available due to the limitations on the bandwidth of video relative to network capacity. A set of quality-sensitive features are extracted from the original video and these features are transmitted to the end users where they are compared with a similar set of features extracted from the test video in order to evaluate video quality.

In NR-VQA methods, there is no access to the reference video or its features. In this case the test video alone is used to estimate the video quality. Some statistical information of the appearance of video or specific artifacts are extracted from the videos [12]. For example sharpness/blurrines or image coding artifacts such as blockiness may be assessed in NR-VQA methods.

In this research, we focus on investigating the impact of frame rate change on video quality, and assume that the original video is available. Therefore, the main focus here is on FR-VQA. Cross-frame rate NR-VQA is an even more challenging problem for future research.

Numerous VQA methods have been proposed and some well-known ones have been made publicly available [38]. Traditional metrics like MSE and PSNR, are still widely used in IQA and VQA applications because they are easy to understand and easy to calculate. SSIM was extended for videos and is widely used in the industry [39]. In practice, the SSIM score of each frame is often calculated and averaged to yield the overall score for the video under test. MS-SSIM is an extension of SSIM that is originally proposed for IQA [23] but has been extended to VQA by averaging or weighted averaging MS-SSIM scores of all frames [38]. Another video-SSIM method [40] uses features of human visual speed perception [41] in its objective model, which are combined with SSIM to calculate the final quality score of video. A VQA method proposed by National Telecommunications and Information Administration (NTIA), named Video Quality Metric (VQM) [42], works with three dimensional spatio-temproal patches and extracts empirical features in both frame and time direction. In the time direction, VQM simply uses the difference of frames. Motion-based Video Integrity Evaluation (MOVIE), the VQA index proposed in [43, 44], not only evaluates quality of videos in space, but also looks at the motion trajectories in video in spatio-temporal space, so as to take into consideration the quality of motion in the overall quality evaluation.

2.3 Frame Rate and Video Quality

A video is a set of images captured in a sequence of consecutive time stamps. Each image in this set is called a frame. Frame rate is the number of frames that are displayed in one second during the video playback and is calculated by the fps unit. Frame rate is an important factor that has direct impact on the perception of motion. Significant effort has been made to explore the values of frame rate with a long history from the first years in the development of cinema to state-of-the-art high frame rate video acquisition and display develiped in the past few years.

In the early 20th century, cameras were capturing videos with frame rates between 16 to 24 fps. By the late 1920's the frame rate of 24 fps became the standard for displaying videos in cinema. This standard has been used in cinemas world wide for many years and is currently still in use. This rate was claimed to give humans the perception of motion without any significant flickering. For different television broadcasting technology, this standard frame rate varies. For European television broadcasting standard which is known as PAL, 23.976 fps (24x1000/1001) is used as the standard frame rate. 25 fps is used in NTSC standard. From the traditional TV systems to the current digital world, there are many display technologies and compression standards, and different frame rates are selected. Examples of the most commonly used frame rates include 23.976, 24, 25, 29.97, 30, 59.94, 60, 120 fps.

The reason for selecting different frame rates depends on the applications and technology constraints. Low frame rates are useful for applications such as video conferencing over mobile devices which desires real-time low data transfer rate, while high frame rates are preferred for digitally stored videos displayed on large screens, digital home cinema, and game entertainment containing high motion contents. The most common frame rates for video broadcast at the moment are 24 and 30 fps. These are classified as low frame rates and their performance is analyzed in [45]. With the advancements in technology and with the increasing availability of high frame rate (60fps, 120 fps) on higher video acquisition and display devices, it is now possible to move towards high frame rate content production and distribution. However, video quality research is largely lacking for videos at frame rates beyond 30 fps, making it difficult to fully understand or justify the benefits of switching to high frame rates.

2.3.1 Frame Rate and Motion

A video may be considered as a three-dimensional signal with two spatial dimensions and one temporal dimension. Frame rate is the temporal sampling rate. With regard to the scene that video is captured from, the frame rate is the number of times that we capture an image from the scene in each second. If there is a moving object in the view of camera, the frame rate is the number of positions of the object that are captured in one second during its movement.

Conversion of high frame rate to low frame rate content is usually performed by dropping certain number of frames. As a three dimensional signal, if this sampling rate (frame rate) is not high enough, aliasing may occur due to the violation of the Nyquist criterion of sampling. Perceptually this may lead to annoying temporal artifacts in motion perception.

Changing frame rate for a video that would be subsequently compressed has more implications than just sampling rate changes. Almost all standard video compression algorithms use motion estimation/compensation of consecutive frames. The motion vectors obtained from compression algorithm is an estimation of motion in the time between two frames. Given a fixed scene, the more time duration between two frames, the larger the motion vector obtained from motion estimation algorithm. With the increase in the time interval between two consecutive frames, the motion vectors will be larger but less accurate. As a result, there would be more residual energy between the real frame and the motion compensated one from motion vectors. Subsequently, more bits are needed to encode the video, and at the same rate of data for each frame, the accuracy would be less and the quality would be degraded.

On the other hand, increase in temporal resolution of video (increase in frame rate) may not result in a monotonic increase of perceived quality because the human visual system may or may not be able to fully capture the fine details contained in high frame rate representations.

In [46] the results of [41] is used for speed sensitivity analysis of human visual perception. This model can be used as an importance factor (of information content) of different regions of a frame. The improvement of quality estimation using the model on top of SSIM is significant while an improvement against PSNR is less obvious. This may be due to the poor quality prediction of PSNR on individual frames.

2.3.2 Frame Rate and VQA Model

Most VQA methods ignore the impact of frame rate. Only a limited number of models consider frame rate as a factor. In [47] a model is proposed based on 5 parameters which include: encoder type, video content, bit-rate, frame rate, and frame size. The low bit-rates of up to 384 Kilo bits per second (kbps) are used with small sizes of the frames QCIF (176×144) and CIF (352×288). Frame rates of 7.5 to 30fps are used. Perhaps the most interesting observation in this research is that in low bit-rate conditions, small frame size is preferred to low frame rate. Ou, *et al.* in [48] performed an analysis on the impact of the frame rate on quality. They have conducted a subjective test with two different resolutions of CIF and QCIF, respectively. They used source videos of 30 fps and created three other frame rates of 15, 10, 7.5 fps by temporal downsampling the source video. The 30fps videos are used as the reference and the MOS obtained from users is analyzed. It is shown that the impact of frame rate on quality can be modeled by using a combination of two exponential terms. In [49] the work is extended and the impact of both frame rate and the quantization parameter is analyzed. An objective model based on the frame rate and quantization parameter is presented that fits the results of the subjective test. The model is composed of a multiplication of two independent functions, one for frame rate and the other for the quantization parameter. In [50], Ou used the result of subjective test for different frame rates of 3.75, 7.5, 15, 30 fps in the combination of 4 different values of Quantizatin Parameter (QP). The spatial term of the model is replaced by a temporal factor and the quantization factor is accounted for by a sigmoid function of PSNR. The resulting method is named Q-STAR. In [51], frame rate and resolution changes are used to estimate video quality. The Spatial Information (SI) is used in combination with the Temporal Information (TI) defined in [15] to create a quality model. A nonlinear model for VQA is proposed in [52] with a combination of frame rate, bit rate, display size, and video content. Six different video contents are coded in different bit rates and CIF display are used. The frame rate is up to 30 fps, and the final model is a nonlinear combination of four different variables corresponding to the four parameters mentioned above.

Low bit rate analysis has been done for different frame rates [53, 54, 55, 56]. In most cases, the low resolution video frames are used in low bitrate videos. The frame rate has been analyzed as a parameter in video broadcasting applications to control the quality of video in transmission over a network [53, 57].

Only in a few works, frame rates above 30fps have been considered. The frame rates of 5, 7.5, 15, 30, 60 in a gaming environment have been analyzed in [58]. It is reported that the higher frame rates better entertain the user and improve the performance of the players. The frame rates above 30fps also have been used in [59, 60] for 3D video Quality of Experience (QoE) assessment. Specific indoor environment with a limited number of subjects has been used for the subjective test.

A cross-frame rate video quality assessment model named FRQM [61] is proposed based on the comparison of temporal wavelet decomposition of a low frame rate video with its original high frame rate version. Spatiotemporal pooling is used to aggregate the comparison results and calculate the overall quality score of each video. For performance evaluation, a cross-frame rate video database consisting 88 videos (22 pristine 120 fps and converted lower frame rates at 15, 30, 60 fps versions) is used. The results show that temporal wavelet decomposition as is used in FRQM is a promising tool in predicting video artifacts in cross-frame rate videos.

VQA of compressed videos with cross-frame rate support has been investigated by learning a model to fuse well-known existing VQA methods [62]. This fusion-based model, named by Video Multi-method Assessment Fusion or VMAF, has been used to predict video quality in different applications such as transcoding and streaming [63, 64, 65, 66].

Considering the efforts already made for VQA of frame rates higher than 30 fps, it is clear that a more comprehensive analysis is desirable for high frame rate VQA. Given that high frame rate content and displays are becoming increasingly popular, it is desired to develop VQA models that appropriately account for the impact of frame rate and use such models to optimize practical video acqusition, compression, transmission, and display systems.

2.3.3 Frame Rate and Rate Control

In many applications such as scalable video coding, there are different parameters that can be adjusted in order to reach lower bit rate, including frame size, compression rate, and frame rate. The goal is to make video bit rate matched with the transmission channel capacity while keeping the video quality as high as possible.

Downsampling the frame size and then interpolating the frame spatially at the user end is one of the methods that is used for scalable video transmission, but this method also causes degradation in quality. Another control factor is the QP value that directly adjust the quantization level in compression. Increasing QP means increasing the quantization step, and as a result leading to a lower bit rate and lower quality. This is a common approach to control the bit rate.

Frame dropping or decreasing the frame rate is another way to control the bit rate. However, converting the frame rate may cause non smooth motion. In addition, receivers should be informed about the altered frame rate for proper playback. Changing the frame rate does not necessarily result in proportional reduction in bit rate. For example if the frame rate of a video is reduced from 30 fps to 15 fps, it does not guarantee that the bandwidth of the new video would be half of the original one. This is because frame dropping affects the accuracy of motion estimation, leading to larger motion compensation residuals to be encoded, and thus more bits are needed for each frame than before in order to maintain similar quality level.

The complications discussed above suggest that using frame rate to control bit rate should be used in a smart way and with caution. For each video, the best solution should be obtained by jointly optimize the bit rate control factors with the guidance of a trusted comprehensive VQA model.

2.4 Evaluation Criteria

An important step in the evaluation of objective VQA models is to validate them using subjective-rated video quality databases, which consist of a collection of videos and their subjective ratings typically in the form of the MOS. Statistical evaluation criteria may be used to compare quality prediction by objective VQA methods against the MOS values obtained from human subjective study. A list of common evaluation criteria used in the literature are as follows.

- Mean Absolute Error (MAE) is defined as

$$MAE = \frac{1}{N} \sum_{i} |S(i) - \hat{S}(i)|, \tag{2.1}$$

  where $S(i)$ is the MOS of the i-th video in the database, and $\hat{S}(i)$ is the score generated by the objective model for the i-th video, and N is the number of videos in the database. Before MAE is computed, a monotonic nonlinear mapping should be used to linearize the mapping and unify the scale between the objective and subjective scores.

- Root Mean Square Error (RMSE) is given by

$$RMSE = \sqrt{\frac{1}{N}\sum_i (S(i) - \hat{S}(i))^2}. \tag{2.2}$$

Similar to MAE, RMSE should be computed after the score mapping between objective and subjective scores.

- Pearson Linear Correlation Coefficient (PLCC) is defined as

$$\begin{aligned} PLCC &= \frac{\sigma_{S\hat{S}}}{\sigma_S \sigma_{\hat{S}}} \\ &= \frac{\sum_{i=1}^{n}(S(i) - \mu_S)(\hat{S}(i) - \mu_{\hat{S}})}{\sqrt{\sum_{i=1}^{n}(S(i) - \mu_S)\sum_{i=1}^{n}(\hat{S}(i) - \mu_{\hat{S}})}}, \end{aligned} \tag{2.3}$$

where, μ_x, σ_x, and σ_{xy} denote the mean and standard deviation of x and the cross-correlation of x and y, respectively. SImilar to MAE and RMSE, PLCC should be computed after the score mapping between objective and subjective scores.

- Kendall's Rank Correlation Coefficient (KRCC) is defined based on the score rank of the test videos in the database. Considering a set of observations of the MOS values and objective scores calculated for videos in a video dataset as (MOS_1, O_1), (MOS_2, O_2), ...,(MOS_n, O_n), the KRCC is given by

$$KRCC = \frac{N_c - N_d}{n(n-1)/2} \tag{2.4}$$

where, N_c is the number of concordant pairs of videos and N_d is the number of discordant pairs in the ranking list of videos under test. A concordant pair is a pair of videos with the same relative order for both MOS values and objective score values in their corresponding ranked order. In another word, two videos (i and j) in a rank order of videos are concordant if their MOS and objective scores follow one of the below two conditions

$$MOS_i < MOS_j \; and \; O_i < O_j \tag{2.5}$$

$$MOS_i > MOS_j \; and \; O_i > O_j \tag{2.6}$$

If a pair of videos are not concordant, they are counted as a discordant pair.

- Spearman's Rank Correlation Coefficient (SRCC) is PLCC between rank values

$$SRCC = PLCC(r, \hat{r}) = \frac{\sigma_{r,\hat{r}}}{\sigma_r \sigma_{\hat{r}}}, \tag{2.7}$$

where r and $\hat{r}$ are the ranks of a video in terms of MOS and objective quality score in the test video database, respectively. Both KRCC and SRCC depend on the score ranks of the test videos in the database only, and are independent from the potential monotonic non-linear mapping that may be applied before they are calculated.

2.5 Summary

In this chapter, we first provided a general overview of IQA and VQA, and discussed some of the well-known IQA and VQA models. We then focused on the frame rate aspect of VQA and discussed how frame rate is related to motion perception, video quality, VQA models, and rate control in video compression. Finally, we reviewed the evaluation criteria of objective VQA models when using subject-rated video databases as benchmarks. Overall, existing studies on the impact of frame rate on video quality are very limited, and advanced VQA models that well account for the impact are largely laking but highly desirable. This inspires us to proceed with deeper investigations of the problem.

Chapter 3

Database Construction and Subjective Study for Cross Frame Rate Video Quality Assessment

The quality of video is ultimately judged by humans. The best objective Video Quality Assessment (VQA) models are those that have the highest correlations with human opinions. Testing VQA methods needs test videos with quality scores given by human subjects. Therefore we create two databases namely Image and Vision Computing lab High Frame Rate Video Quality Assessment video database one (IVC-HFRVQA-I) and Image and Vision Computing lab High Frame Rate Video Quality Assessment video database second (IVC-HFRVQA-II). We also conducted two subjective studies on the databases. Using the data collected through the IVC-HFRVQA-I and IVC-HFRVQA-II databases, we evaluate the performance of well-known VQA models.

3.1 IVC-HFRVQA-I Video Database

3.1.1 Database Construction

A new database of videos is built to investigate the impact of frame rate on top of 7 pristine videos. Videos are selected from online resources of videos on the Internet including technical YouTube channels. In the selection process, we tried to cover diverse content and motion types. In addition, the perceptual quality of the source videos and the coverage of different combinations of spatial and motion complexities were considered. The source pristine videos all have a frame rate of 60 fps. All the videos are 10-second long and with a resolution of 1920×1080 in YUV420 color format. Sample frames of the source videos are shown in Figure 3.1. The specification of the content types of these source videos used to generate the IVC-HFRVQA-I video database is summarized in terms of "object motion", "camera motion", and "spatial complexity" in Table 3.1. The source video is compressed with FFMPEG using H.264 compression standard. Different configurations of compression, resolution, and frame rate are used to generate the comprehensive database. The values used for different parameters in generating this database are summarized in Table 3.2. We used the following command template to encode videos using FFMPEG.

```
ffmpeg -i input-file -c:v libx264 -crf qp -r fr -s res output-file
```

where *input-file* and *output-file* are the address of the input and output files respectively, and *qp*, *fr*, and *res* represent the compression level (qp), frame rate, and spatial resolution, respectively.

Two different spatial resolutions are used: 640×480 progressive scan (480p) and 1920×1080 progressive scan (1080p). 480p represents standard definition (SD) formats and 1080p is

Table 3.1: Details of the source videos in IVC-HFRVQA-I database

Sequence	Obj. Motion	Camera Motion	Spatial Complexity
Battle	High	Yes	High
Beach	High	Yes	Low
Carousel	Medium	No	High
Notre Dame	Medium	Yes	High
Guys	High	No	Low
Sea	Low	Yes	Low
Talk	Low	No	Low

Table 3.2: Configurations used to generate test videos from the source video

Parameter	Values
Frame Rate	5,10,15,30,45,60
Quantization Level	22, 27, 32, 37
Frame Size	$640 \times 480, 1920 \times 1080$

a common High Definition (HD) format supported by all HDTV display devices. The 480p videos are generated from 1080p original videos by bi-cubic interpolation followed by down-sampling. Different frame rates from 5fps to 60fps were generated for different combinations of quality, resolution, and content. The values of frame rate have been selected based on different needs. 30 fps is a common frame rate in many current applications. 15fps and the lower frame rates are often used to support lower bit-rate encoding as a compromise for limited storage space or transmission bandwidth. 60fps is the most common high frame rate being used in practice. 45fps is the middle frame rate that is included to make a better spaced temporal resolution in the subjective test. Different frame rate has been generated using FFmpeg tool using dropping and duplicating methods. Four different QP values have been used in order to cover different levels of compression from low compression of QP=22 to high compression of QP=37. As such, for each source content, there are 6(Frame rate) x 4(QP) x 2(Resolution) = 48 test video sequences. Altogether, there are totally 48 x 7 = 336 video sequences in the database.

There are three important features of the database. First, the database contains sequences with a wide range of frame rates from 5fps to 60fps, which allows us to directly examine the general trend of the impact of frame rate on perceptual video quality. The better coverage of the frame rates makes the database better suited to study a wider range of practical applications, and to better observe the general trend of quality variations as a function of frame rate that could be extrapolated beyond the frame rates currently being tested. Second, the database contains sequences with different combinations of spatial complexity, object motion, and camera motion, allowing us to study the interactions between frame rate and video content. Third, the database contains sequences with different compression levels and frame sizes, allowing us to investigate the trade-offs between frame rate, compression level, and spatial resolution.

Compared with the new database, existing databases in the literature are limited in one aspect or another. In [17], the authors attempted to consider time complexity with motion, but only videos with low spatial resolution (352x240) and frame rates (up to 30fps) were used. Similarly, in [2, 3, 5, 4, 18], only small resolution videos (CIF or QCIF size) were employed. In [9] only low bit rate videos are considered, which are not able to cover the HD cases where the bit rates are often much higher. In [14, 15], 60fps videos were studied, but the impact of spatial and temporal complexities on video quality was investigated separately, making it impossible to study the combined effect of complexities as well as variations in video content and quantization levels. In [19], the effects of quantization and frame rate were studied while the dimension of spatial resolution and content complexity were missing, making it difficult to build or test a complete model.

Table 3.3: Parameters of display and viewing conditions in subjective study

Parameter	Value
Subjects Per Monitor	1
Screen Resolution	1920 × 1080
Screen Diameter	31.5”
Viewing Distance	30.00”
Screen Width	27.45”
Viewing Angle	49.2°H/28.9°V
Screen Height	15.44”
Pixels Per Degree	78.1/74.8 pixels(H/V)

3.1.2 Subjective Study

A subjective study was done to obtain human subjective quality scores of the videos in IVC-HFRVQA-I database. 25 people including 13 female and 12 male aged between 22 to 33 scored the quality of all 336 videos in the database. The subjective test was done in Image and Vision Computing at University of Waterloo. The subjects asked to score the quality of each video based on their overall perception of quality. Each sequence is displayed to each subject once. The floor, ceiling, and walls of the experiment room had no reflection and was not insulated by any audio/visual pollution. The display and viewing conditions used in the subjective test are shown in Table 3.3.

A single stimulus, 11-grade numerical categorical scale (SSNCS) protocol was employed in this subjective test. A general introduction was given at the beginning of the whole test, and more specific instructions and a training session were given afterwards. The video content of the training videos is similar but different from those in the formal test session. The parameters used to generate the training videos are also similar to the test video parameters. The subjects were asked to rate training videos until they fully understood the requirements and stabilized their rating strategies.

All stimuli were displayed in actual pixels, and in the case of 480p sequences, display regions outside the frames were filled with black pixels. A still gray image was displayed for 7 seconds after each test video for subject scoring. Each stimulus was shown once and the order of stimuli was randomized. Eighty-four videos were evaluated in one session. To reduce visual fatigue, each session was controlled to be within 20 minutes and sufficient relaxation periods (5 minutes or more) were given between sessions. The MOS for each test video was computed using scores of all users. In the next section we focus on the impacts of frame rate on perceived video quality with different quantization levels, different frame sizes, and different complexities of spatial content and motion.

3.1.3 Key Observations on Subjective Study Results

Based on the subjective test results, we have carried out a series of statistical analysis. In this section we focus on a few main observations that are related to the design of our cross-frame rate objective VQA models later.

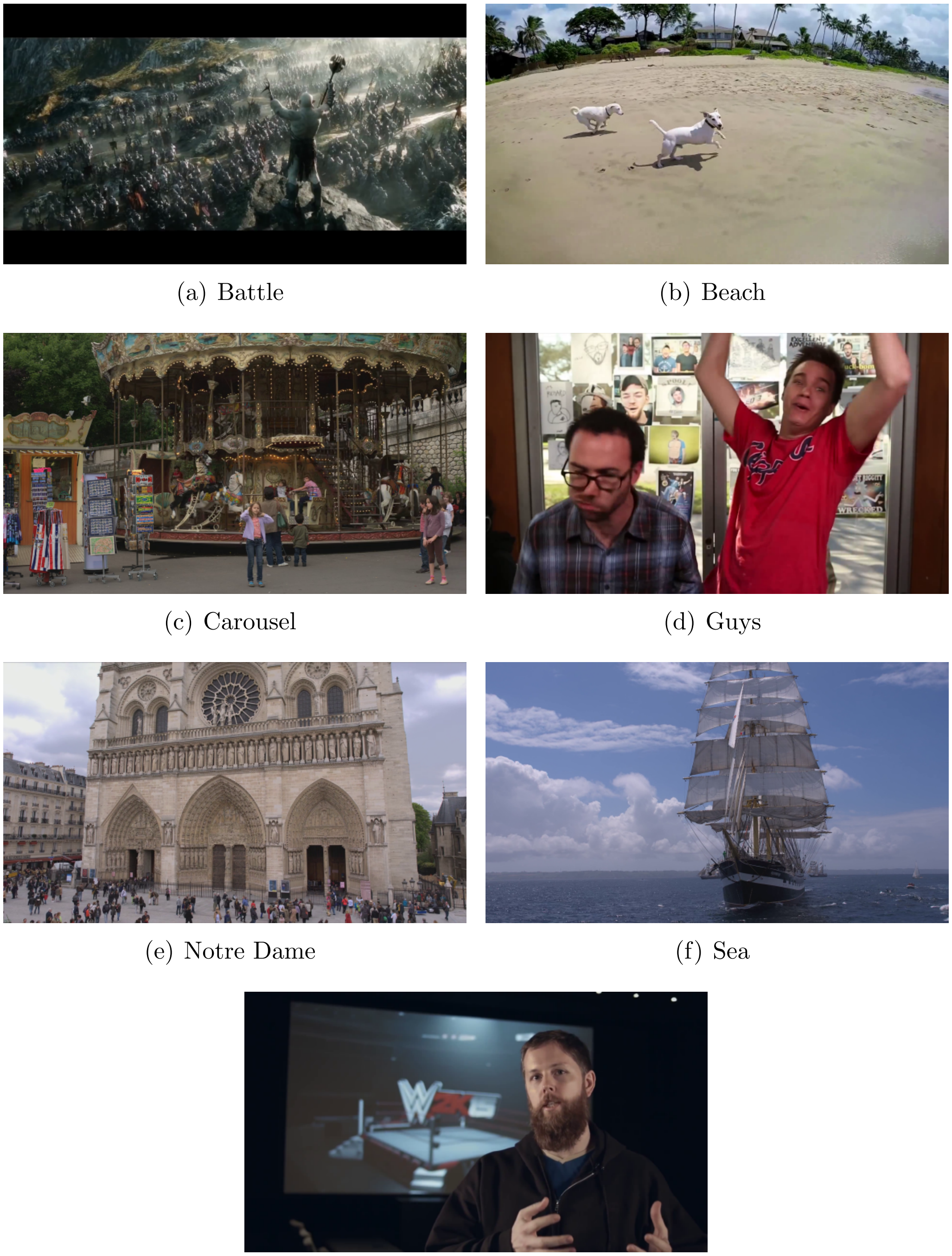
Fig. 3.2 shows the MOS values for all source sequences with respect to different quantization levels (QP values) and different frame sizes (480p or 1080p). It can be observed that there is a significant improvement in terms of MOS values from 5fps to 30fps, which is consistent with previous results [50, 60]. Such improvement decreases with increasing frame rate, especially after 30fps. Even though small, the improvement from 30fps to 60fps can still be clearly discerned, which justifies the value of going beyond 30fps. The general trend being observed here suggests that the quality improvement saturates at high frame rates, thus increasing frame rate beyond 60fps may not lead to distinguishable quality gain, depending on video content. Scrupulous observers may find that the improvement from 30fps to 45fps seems to be below expectation from the general trend. This may be because unlike 5fps, 10fps, 15fps, and 30 fps videos, the 45fps videos could not be generated directly by uniformly picking one of every integer number of frames from the source video sequences of 60fps. Instead, three of every four frames were picked, which affects the uniformity of frame time-spacing. An alternative way of creating 45fps video from 60fps ones is to temporally interpolate and insert new frames to satisfy the uniform time-spacing condition. However, the interpolation process creates additional quality degradations of the video.

Across distortion levels, it can be seen that the quality improvement decreases with the level of quantization, where QP = 22 (less compression, higher quality) shows the largest improvement and QP = 37 shows the least improvement. This implies that there is a competing relationship in terms of perceived video quality between reducing compression artifact and increasing frame rate. Previous work [48, 49] addressed this aspects for 5fps to 30fps videos and proposed certain computational VQA models to compromise both factors. However, this trend saturates again in the range of 30fps to 60fps, which indicates that previously developed models need to be reexamined for their generalization ability to high frame rate levels.

Fig. 3.3,3.4 reports the MOS values for different complexity levels of spatial content with respect to quantization level and spatial resolution. A similar general trend of quality versus frame rate is observed. An interesting point to notice here is that for the case of 480p videos, although the MOS curves corresponding to low and high spatial complexity videos almost overlap with each other from 5fps to 30fps, there is a significant gap between them from 30fps to 60fps, where low complexity videos always obtain lower MOS values. One potential explanation is that high frequency, high texture complexity videos desire not only higher spatial sampling rate but also higher temporal sampling rate in order to accurately represent the complex content without strong (aliasing) artifacts, especially when there is motion associated with the complex textures. As a result, when the frame rate goes from low to high, humans recognize more quality improvement than that from relatively simple texture content. In the case of 1080p videos, the spatial resolution is already sufficient to precisely represent more complex content, and thus the benefit of moving towards high frame rate is less pronounced.

Fig. 3.5 and 3.6 reports the MOS values for different levels of object motion (low, medium and high) with respect to different quantization levels and different frame sizes. Based on previous studies (e.g. [59, 60]), it was expected that there exists some strong object motion dependency, i.e., with increasing frame rate, higher object motion videos would pronounce more improvements than lower ones. Somewhat surprisingly, this is not the case in our experiment, as no clear object motion dependency can be found in Fig. 3.5. Through more careful observations of the data and discussions with the subjects who did the experiment, we found two possible explanations. First, the uncertainty of human visual perception increases with the speed of motion [46, 67]. When the object motion is extremely high, the perceptual uncertainty becomes so high that further increasing frame rate would not help the visual system to capture more information from the scene. Second, in the case of low to moderate object motion, if they are accompanied by slow camera motion, humans tend to be more sensitive to temporal artifacts [1] and thus the effect of increasing the frame rate could be strong. It is also worth noting that the trend is independent of the quantization level.

The way the new database was built allows us to examine not only the impact of individual parameters including frame rate, quantization level, and spatial resolution on the overall video quality, but also their combined effect in a joint parameter space. Fig 3.7 (a) and Fig 3.7 (b) show the overal MOS score as a joint function of frame rate and quantization level, for 480p and 1080p resolution videos respectively. It can be seen that although increasing frame rate is generally helpful in improving the overall video quality, the speed of improvement depends on the quantization level. In other words, the overall quality improvement is not a simple additive effect of improving frame rate and reducing quantization. Their interactions need to be taken into account. A similar conclusion may be drawn when we include the spatial resolution parameter into the equation. Moreover, the results we presented earlier also show that spatial and motion complexities are adding more complications into the picture. Therefore, building a comprehensive objective quality prediction model that considers the impact of all parameters is a challenging but important task that desires deeper understanding and further investigation.



(a) Battle

(b) Beach

(c) Carousel

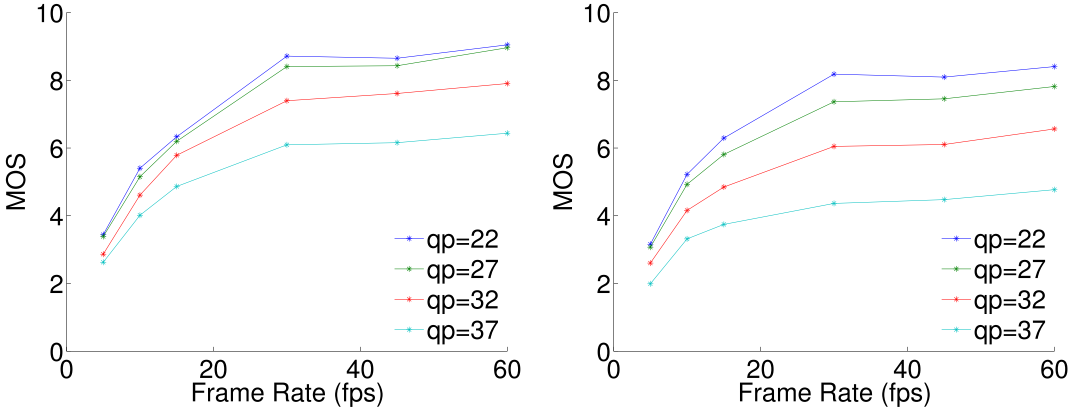
(d) Guys

(e) Notre Dame

(f) Sea

(g) Talk

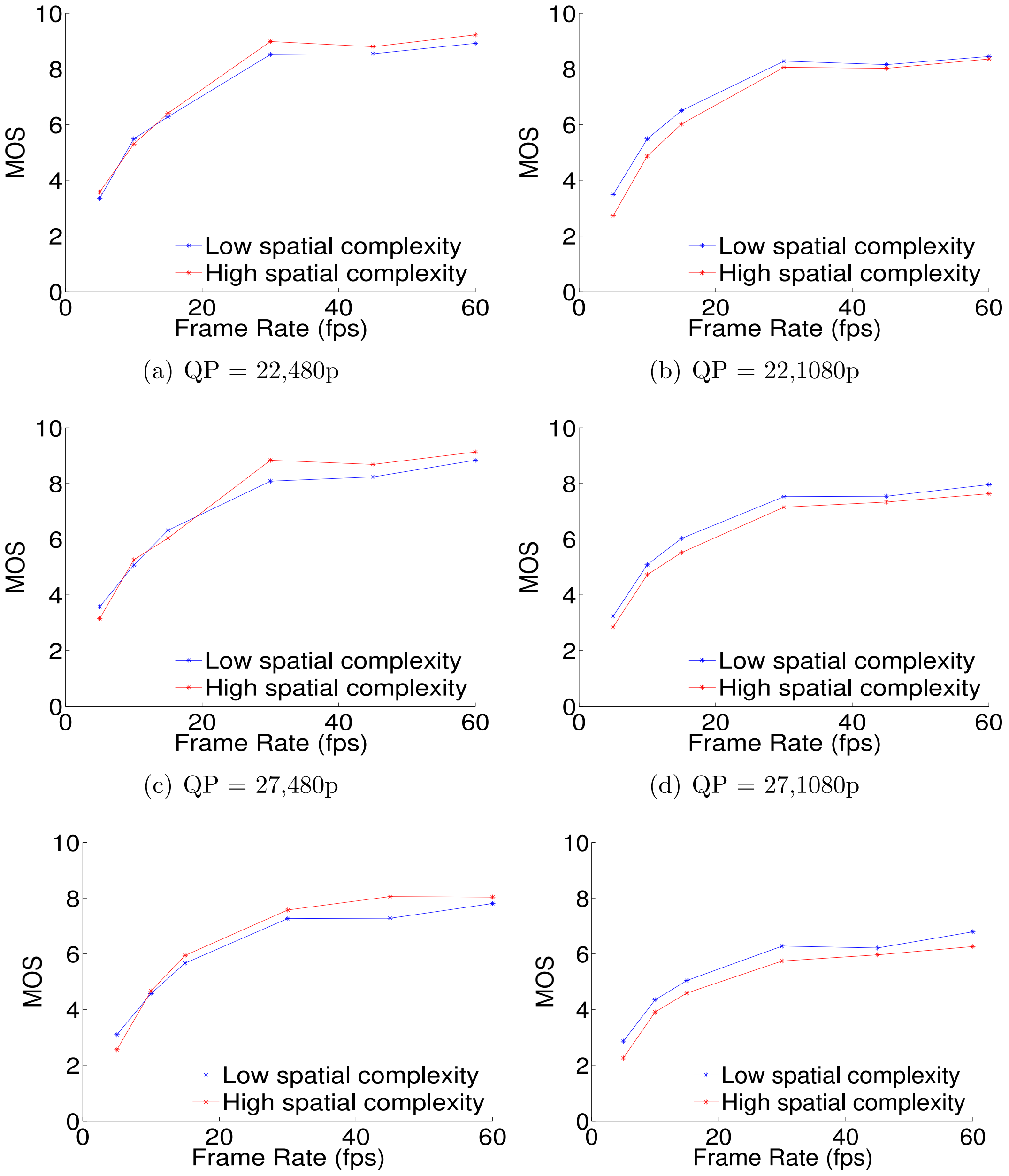
Figure 3.1: Sample frames of the source pristine videos used in the database.



(a) 480p

(b) 1080p

Figure 3.2: MOS versus frame rate for all test videos.



(a) QP = 22,480p

(b) QP = 22,1080p

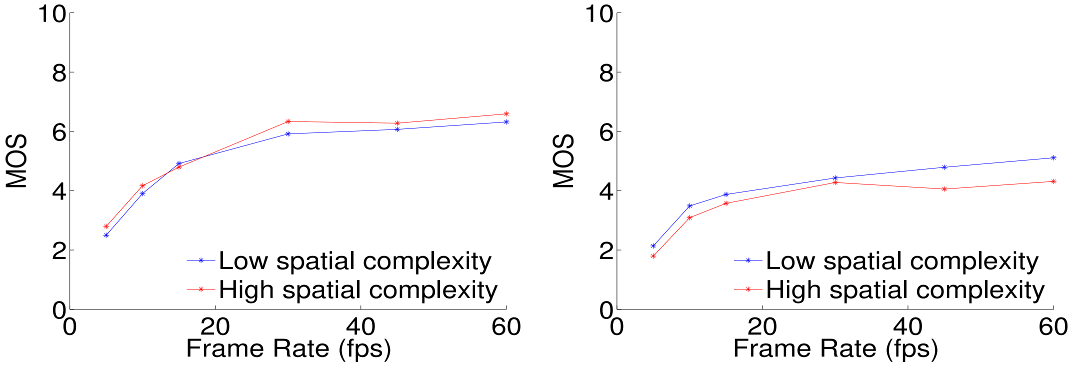
(c) QP = 27,480p

(d) QP = 27,1080p

(e) QP = 32,480p

(f) QP = 32,1080p

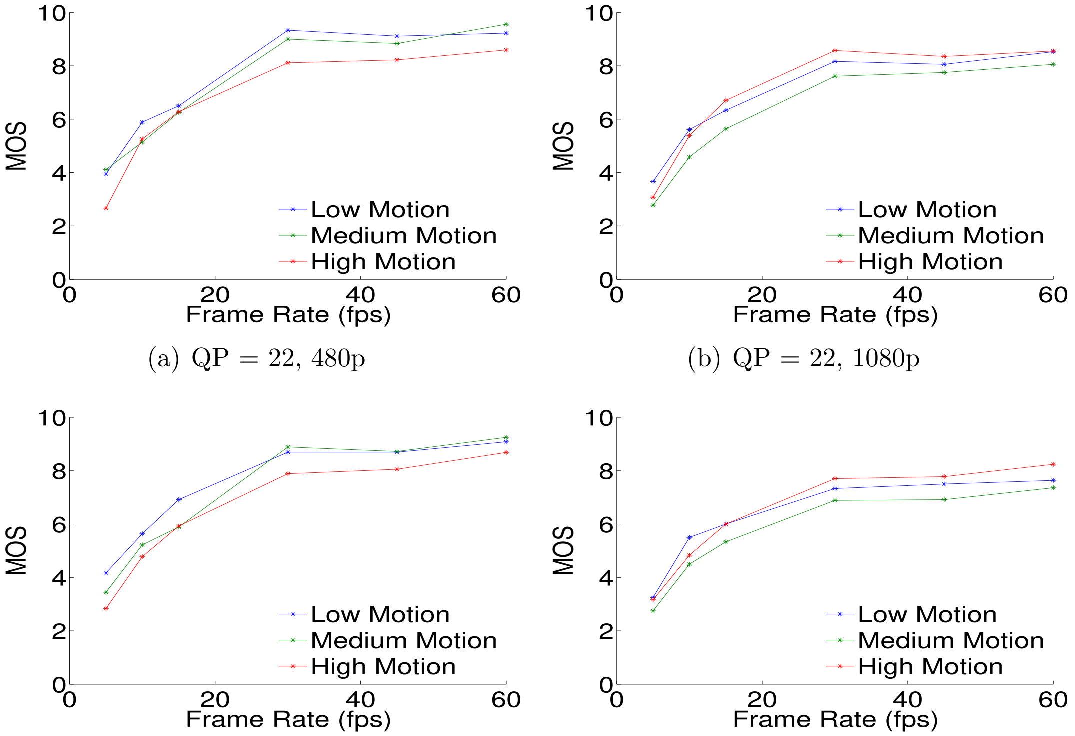
Figure 3.3: MOS versus frame rate for videos with low and high spatial complexities.



(a) QP = 37,480p

(b) QP = 37,1080p

Figure 3.4: MOS versus frame rate for videos with low and high spatial complexities(cont'd).



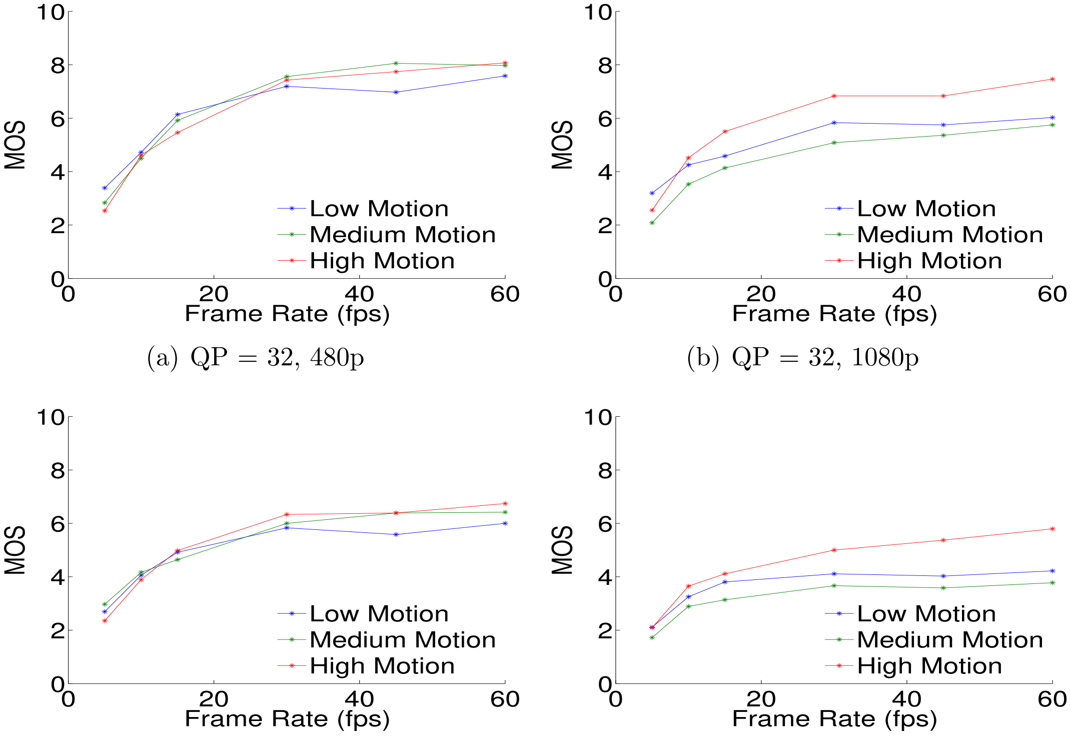
(a) QP = 22, 480p

(b) QP = 22, 1080p

(c) QP = 27, 480p

(d) QP = 27, 1080p

Figure 3.5: MOS versus frame rate for videos with low, medium and high object motion.



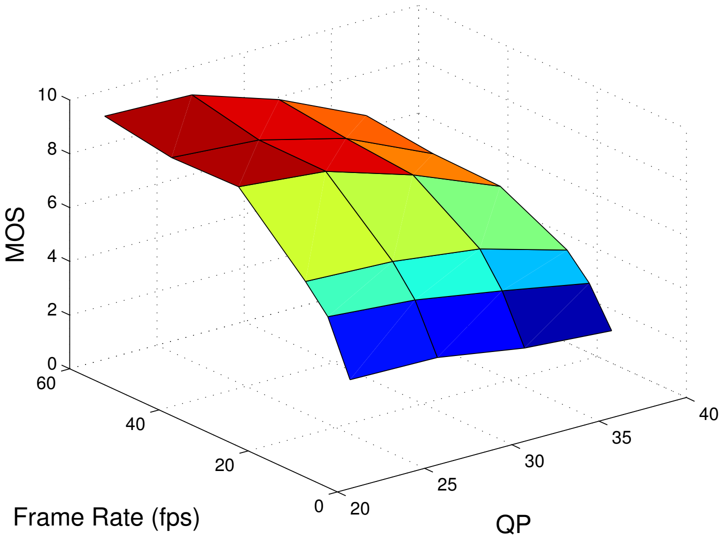
(a) QP = 32, 480p

(b) QP = 32, 1080p

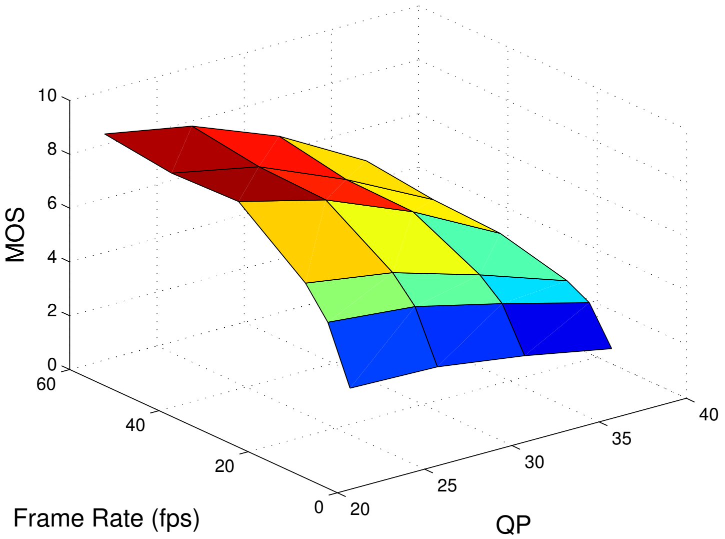
(c) QP = 37, 480p

(d) QP = 37, 1080p

Figure 3.6: MOS versus frame rate for videos with low, medium and high object motion.



(a) 480p



(b) 1080p

Figure 3.7: MOS as a function of frame rate and quantization parameter for 480p (a) and 1080p (b) videos.

3.2 IVC-HFRVQA-II Video Database

3.2.1 Database

The Waterloo-IVC High Frame Rate Video Quality database two (IVC-HFRVQA-II) is constructed using various frame rates, resolutions, and compression quantization levels. IVC-HFRVQA-II is built from 10 pristine source HFR videos from the BVIHFR dataset [68] and spans diverse content, including humans, plants, natural scenes, objects, and synthetic scenes. The source videos are at a resolution of 1920×1080, and a high frame rate of 120 fps. Fig 3.8 shows the screenshots of the video contents in the IVC-HFRVQA-II dataset. The detailed specifications of the contents are listed in Table 3.4. As it can be seen in Table 3.4, the video contents are selected from various combinations of object motion, camera motion, and spatial complexity.

Table 3.4: Specification of pristine videos used in IVC-HFRVQA-II dataset.

Sequence	Object Motion	Camera Motion	Spatial Complexity
Bubblehead	High	No	High
Books	Low	Yes	High
Bouncy ball	Medium	No	Low
Catch	Low	No	High
Catch and Track	Low	Yes	High
Cyclist	Medium	Yes	Medium
Guitar Focus	Low	No	Low
Hamster	High	No	High
Lamppost	Medium	No	High
Plasma	High	No	Low

Table 3.5: Configurations used to generate test videos in IVC-HFRVQA-II database from the source videos.

Parameter	Values
Frame Rate (fps)	15,30,60, 120
Quantization Parameter(QP)	27, 32, 37, 42
Frame Size	320×240,640×480,720×1280, 1920×1080

Using the aforementioned sequences as the source, each video is encoded into four quantization levels with FFMPEG (H.264 compression standard) at 4 spatial resolutions and 4 frame rates to cover diverse quality levels. The choices of resolution are 320×240, 640×480, 1280×720, and 1920×1080, covering a wide range of common video resolutions used in different devices and different networks. Frame rate choices are based on the commonly used parameters for transmission of HD videos over networks. To be specific, 15, 30, 60, and 120 fps correspond to mobile network display, standard display, HFR TV display, and gaming monitor displays, respectively. Table 3.5 summarize the configurations used to generate test videos in IVC-HFRVQA-II database from the source videos. In total, IVC-HFRVQA-II database contains 480 videos. We used the following command template to encode videos using FFMPEG.

```
ffmpeg -i input-file -c:v libx264 -crf qp -r fr -s res output-file
```

where *input-file* and *output-file* are the address of the input and output files, respectively, and *qp*, *fr*, and *res* represent the compression level (qp), frame rate, and spatial resolution, respectively.

In comparison to the IVC-HFRVQA-I database [69], the IVC-HFRVQA II database has a more complete family of resolutions, from a low resolution of 320×240 to a high resolution of 1920×1080, while in IVC-HFRVQA-I there are only two resolutions: one for Standard Definition (SD) and one representing High Definition (HD) resolutions. In addition to the extension to the number of resolutions, a higher frame rate of 120 fps is added to IVC-HFRVQA-II in comparison to IVC-HFRVQA-I as this frame rate has recently become more common in gaming monitors and is proposed for future cinema standard. Similar to IVC-HFRVQA-I, four different quantization levels are used in the IVC-HFRVQA-II database

Table 3.6: Comparison of configuration parameters in cross-frame rate VQA databases.

Database	Spatial Resolutions (pixel)	Temporal Resolutions (fps)
Zhai et al. [47]	CIF, QCIF	30, 15, 7.5
Ou et al. (I) [48]	CIF, QCIF	30, 15, 7.5
Ou et al.(II) [49]	CIF, QCIF	30, 15, 7.5
Janowski [51]	SD, CIF, QCIF, SCIF, SQCIF	30, 25, 20, 15, 10, 5
Banitalebi et al. [59]	Full HD (1080p)	60, 48, 30, 24
Mackin et al. [68]	Full HD (1080p)	120, 60, 48, 30, 24
IVC-HFRVQA-I. [69]	1080p, 480p	60, 45, 30, 15, 10, 5
IVC-HFRVQA-II	1080p, 720p, 480p, 240p	120, 60, 30, 15

in order to cover different levels of compression and have various numbers of compression artifacts. The quantization level varies from a low compression level of QP=27 to a high compression level of QP=42. The comparison of IVC-HFRVQA-I and IVC-HFRVQA-II and other cross-frame rate VQA databases is summarized in Table 3.6

3.2.2 Subjective Study

The subjective testing experiment is set up as a normal indoor home setting with ordinary illumination level, with no reflecting ceiling walls and floors. All videos are displayed at full screen on an LCD monitor at a resolution of 1080 × 1080 pixels with True color (32 bit) at 165 Hz. The monitor is calibrated in accordance with the recommendations of ITU-T BT.500 [70]. A customized graphical user interface is used to render the videos on the screen in random order and to record the individual subject ratings on the database. The details of the display parameters and viewing conditions are reported in Table 3.7. The study adopts a single-stimulus quality scoring strategy. A total of 36 naïve subjects, including 20 males and 16 females aged between 21 and 38, participated in the subjective test. Visual acuity (i.e., Snellen test) and color vision (i.e., Ishihara) are confirmed from each subject before the subjective test. The study took about two hours, which was divided into 4 sessions with 5-minute breaks in between to avoid the fatigue effect. A general introduction was given at the beginning of the test, and more specific instructions and a training session were given after this. The video content of the training videos was similar, but different from those in the formal test session. The parameters used to generate the training videos are also similar to the test video parameters. The choice of a 100-point continuous scale as opposed to a discrete 5-point ITU-R Absolute Category Scale (ACR) has advantages of expanded range, finer distinctions between ratings, and demonstrated prior efficacy [71].

The raw subjective scores are used in the subsequent analysis. After the subjective user study, 5 outliers are removed based on the outlier removal scheme suggested in [70]. The final quality score for each individual video is computed as the average of subjective scores, namely the mean opinion score (MOS), from all valid subjects.

For analysis of the reliability of the MOS values, we evaluated the correlation between the scores given by each participant in our subject study to different videos by the average of quality score of each video. Figure 3.9 shows the Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank correlation Coefficient (SRCC) for each individual's scores with the average quality score measured by MOS. From the results on Figure 3.9, we observe significant general agreement between the participants on scoring of the video quality.

In addition to the evaluation of correlation of each participant's score to videos with the MOS values, we also investigate the variation of scores given to each video with specific compression parameters by different participants. Figure 3.10 shows the average and error bar (±1 std) of scores given by participants to different video sequences for loss-less compression (QP=0) and high spatial resolution (1080p) across different frame rates. We

Table 3.7: Viewing conditions of the subjective test.

Parameter	Value
Subjects Per Monitor	1
Screen Resolution	1920 × 1080
Refresh Rate	165 Hz
Screen Diameter	27”
Viewing Distance	30.00”
Screen Width	23.5”
Screen Height	13.2”
Aspect Ratio	16:9
Viewing Angle	49.2°H/28.9°V
Pixels Per Degree	78.1/74.8 pixels(H/V)

observe that for higher frame rate there is less variation of scores a high quality level. Figure 3.11 also shows the same analysis for videos with loss-less compression (QP=0) and high frame rate (120 fps) and across different resolutions in the subjective study on IVC-HFRVQA-II database. Moverover, the impact of compression on variation of subjects' scores was investigated for videos with high spatial resolution (1080p) and high frame rate (120 fps) in the subjective study on IVC-HFRVQA-II in Figure 3.12.

3.2.3 Key Observations on Subjective Study Results

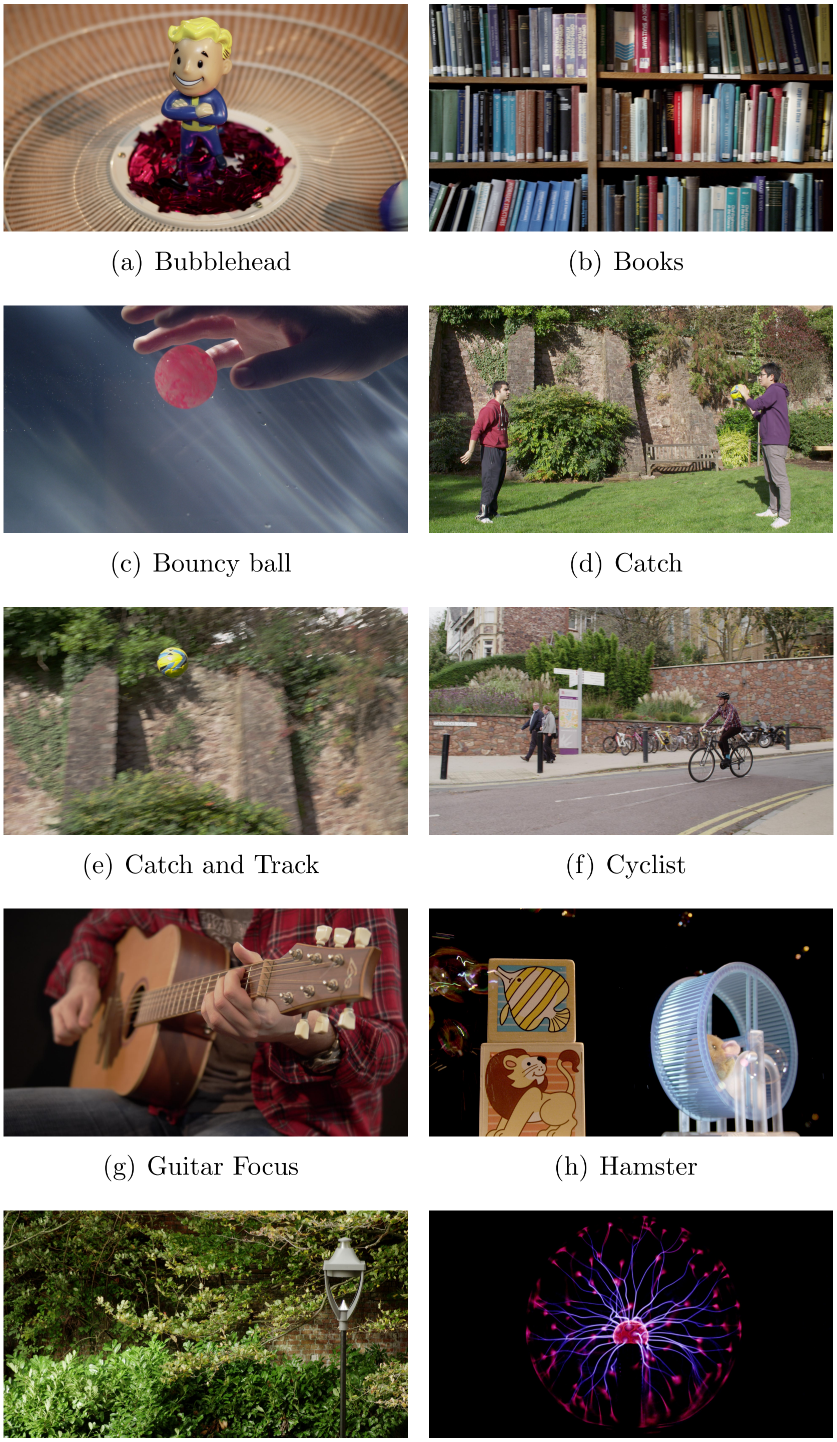
The analysis of the subjective test results is reported in this section. There are many interesting observations that can be discussed regarding the impact of frame rate on video quality. We investigate the average subjective quality scores of videos in IVC-HFRVQA-II with different viewpoints to explore the integrated impact of frame rate and other parameters including resolution and quantization on video quality.

The first observation comes from the overall trend of perceptual quality of video in different frame rates. Fig. 3.13.a shows the MOS values of all source sequences with respect to different frame rates. It can be seen from Fig. 3.13.a that the there is significant improvement in terms of MOS values from 5 fps to 60 fps but such improvement in quality score is not the same with the increase of frame rate to 120 fps. Fig. 3.13.b shows the MOS values of only source videos with lossless compression (QP=0). Comparing part a and b of Fig. 3.13 it is obvious that the improvement in the quality score between 60 fps to 120 fps is more significant for videos without lossy compression. This shows that the impact of frame rate on video quality is not independent of the compression. In fact, the impact of frame rate change on the quality score is more significant for the users when there is no information loss due to compression.

Fig. 3.14 shows the MOS values of all source sequences with respect to different frame rates and different QP values. It can be seen in Fig. 3.14, the difference in the quality trend, when the frame rate increases, is more significant for lower quantization values and for lower bit-rate compression (or equally higher values of QP) the frame rate increase does not lead to an increase in quality. Also, the impact of compression on video quality is always perceivable by human subjects in different frame rates and resolutions.

To analyze the impact of resolution and frame rate on the results of subjective testing, the MOS values of all source sequences with respect to different frame rates and different resolutions are shown in Fig. 3.15.a . Fig. 3.15.b also shows the MOS values with the same grouping of resolutions and frame rates but only for lossless compressed videos (QP=0). According to Fig. 3.15.a, the quality improves with the increase of frame rate for different groups of videos with different resolutions; however, this trend is more significant for higher resolutions. From the observation on the impact of resolution and frame rate on quality, it can be seen that the impact of resolution reduction on video quality is more significant for frame rates higher than 30 fps.

To investigate the impact of frame rate increase on the quality of different source contents in IVC-HFRVQA-II, the MOS values of different source sequences with respect to different frame rates is shown in Fig. 3.16. An interesting observation from Fig. 3.16 is that the quality trend after 30fps varies across content. It reveals the fact that the impact of higher frame rate on video quality is strongly dependent on the content and might not be significant at all for some content. This content dependency should be investigated more in a video dataset with more various content. An effort to investigate this content dependency is done by grouping videos based on different criteria in the current database. For example, we group the videos into two groups: moving camera and still camera based on the existence of camera motion. Fig. 3.17 shows the MOS values of all sources in these two groups with respect to different frame rates. As it can be observed in Fig. 3.17 the impact of frame rate increases on video quality improvement is more significant for the moving camera video group. It can be concluded that frame rate increase can be more beneficial in quality improvement when there is a general motion.



(a) Bubblehead

(b) Books

(c) Bouncy ball

(d) Catch

(e) Catch and Track

(f) Cyclist

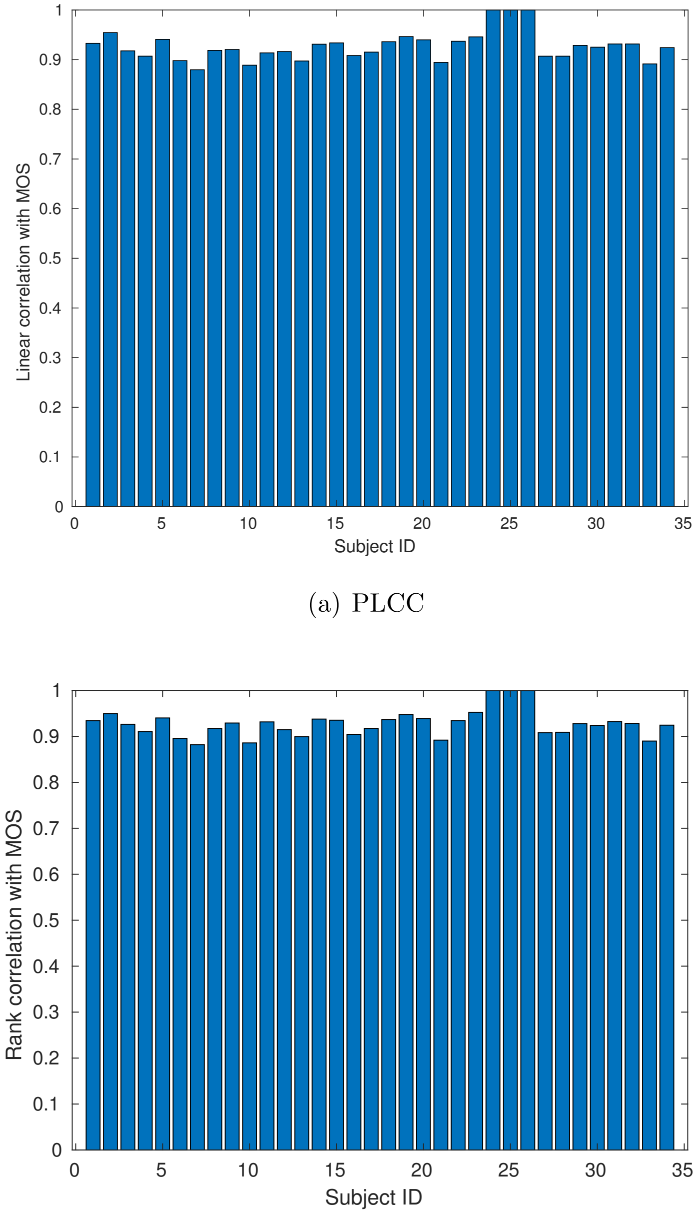
(g) Guitar Focus

(h) Hamster

(i) Lamppost

(j) Plasma

Figure 3.8: Sample frames from the pristine videos used in IVC-HFRVQA II database.



(a) PLCC

(b) SRCC

Figure 3.9: Correlation between each participant's score and MOS values on IVC-HFRVQA-II database.

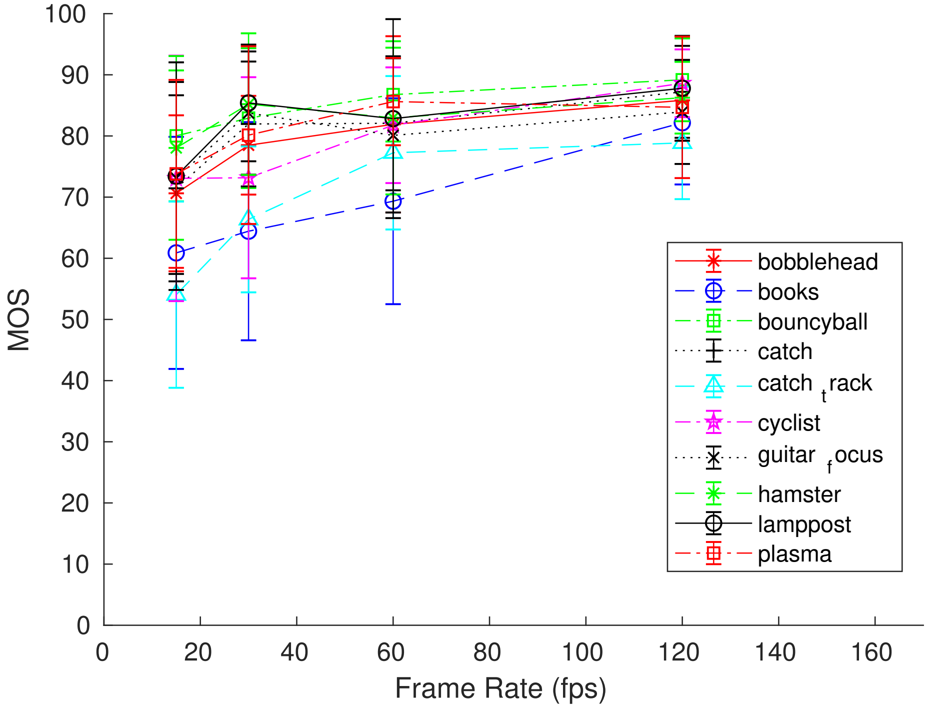


Figure 3.10: Average and error bar (±1 std) of participants' scores to each video sequence with QP=0 and spatial resolution of 1080p across different frame rates on IVC-HFRVQA-II database.

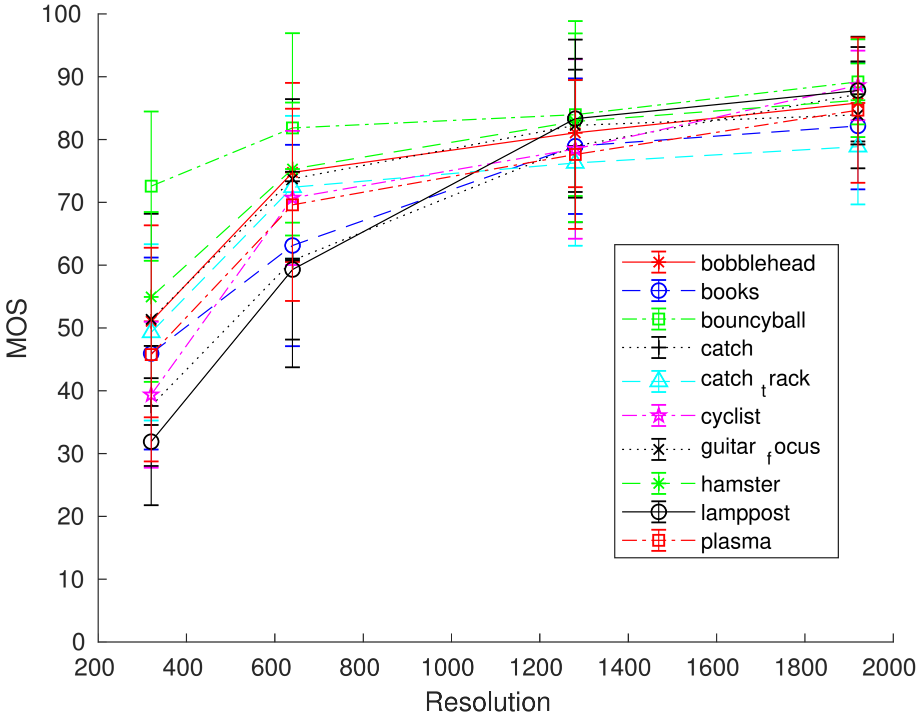


Figure 3.11: Average and error bar (±1 std) of participants' scores to each video sequence with QP=0 and frame rate of 120fps across different resolutions on IVC-HFRVQA-II database.

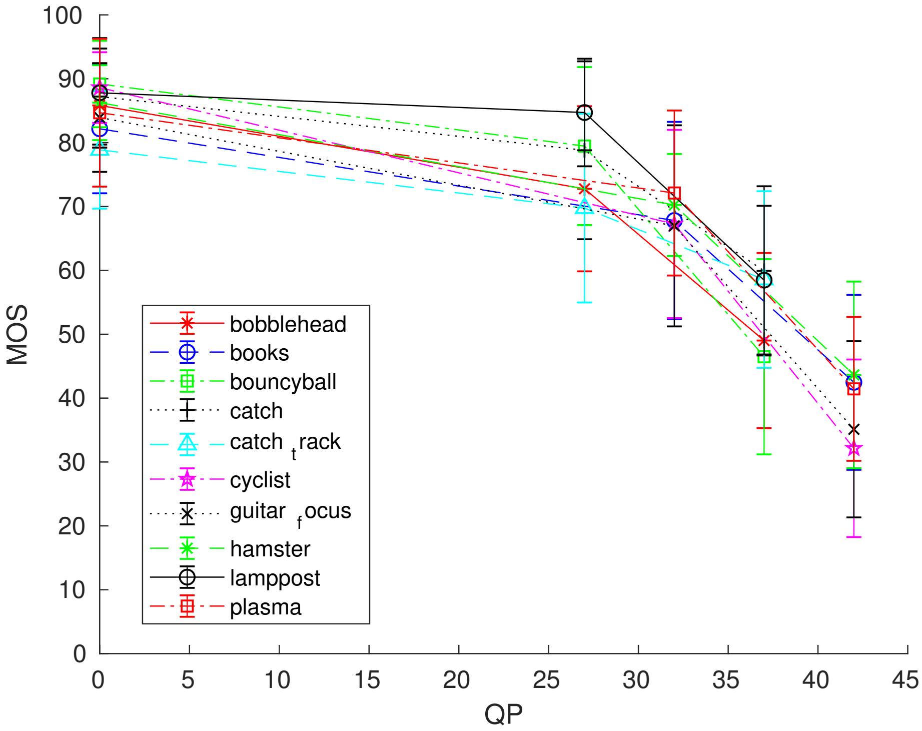
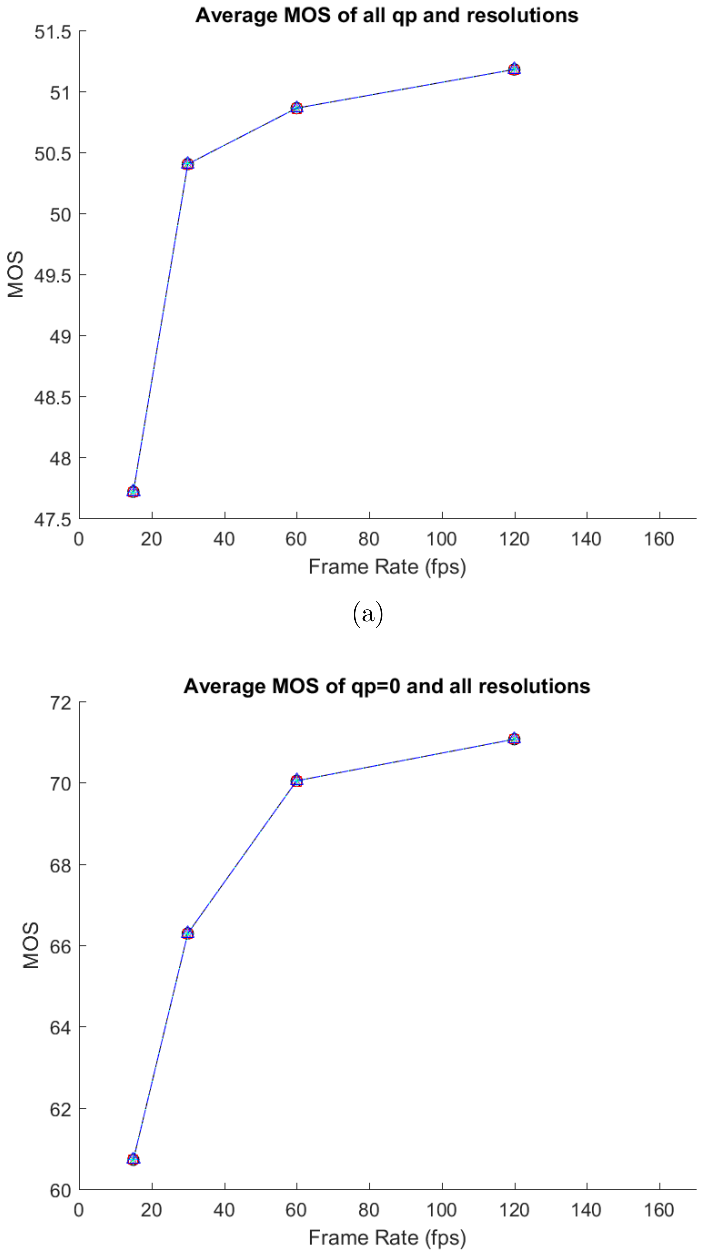


Figure 3.12: Average and error bar (±1 std) of participants' scores to each video sequence with frame rate of 120fps and spatial resolution of 1080p across different compression level (QP values) on IVC-HFRVQA-II database.



(a)

(b)

Figure 3.13: MOS versus frame rate for all test videos in IVC-HFRVQA-II database.

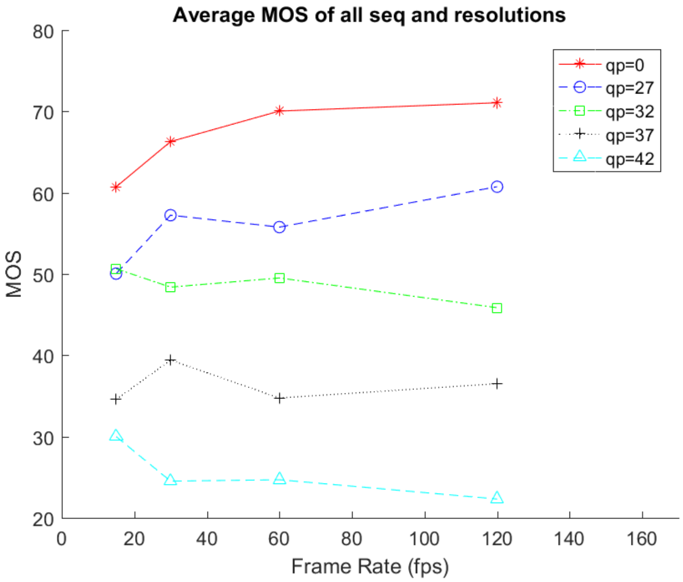
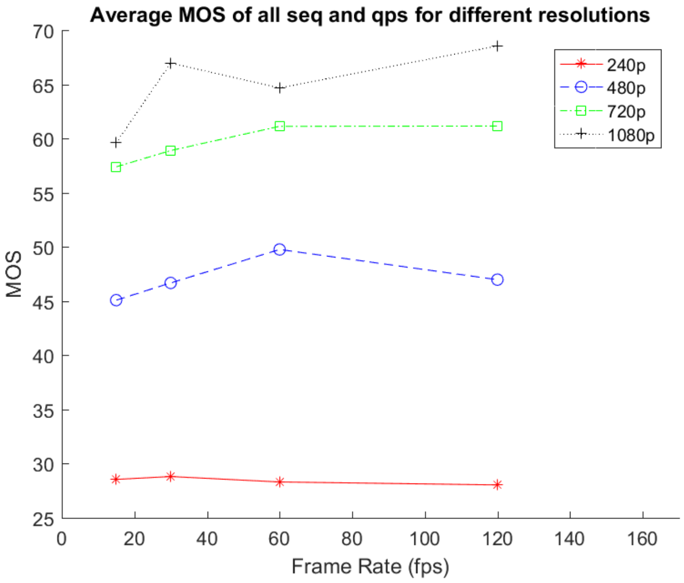
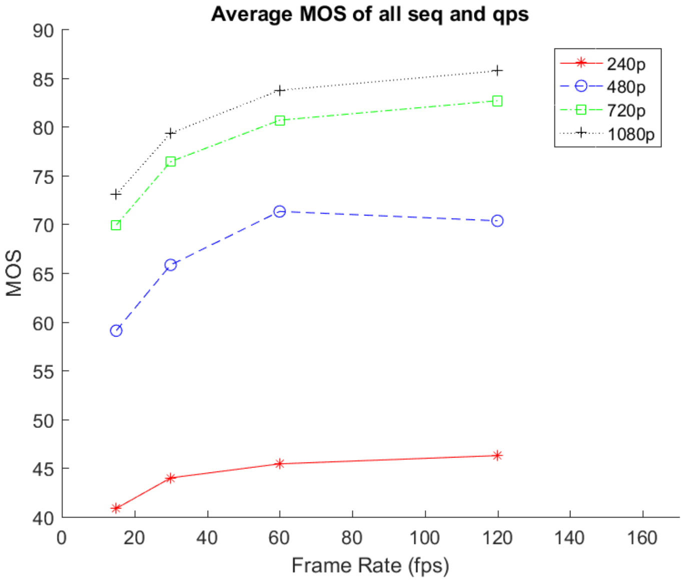


Figure 3.14: MOS versus frame rate for test videos in IVC-HFRVQA-II database grouped by QP values.



(a)



(b)

Figure 3.15: MOS versus frame rate for test videos in IVC-HFRVQA-II database grouped by different resolutions a) for all QP values, b) for only lossless compressed videos (QP=0).

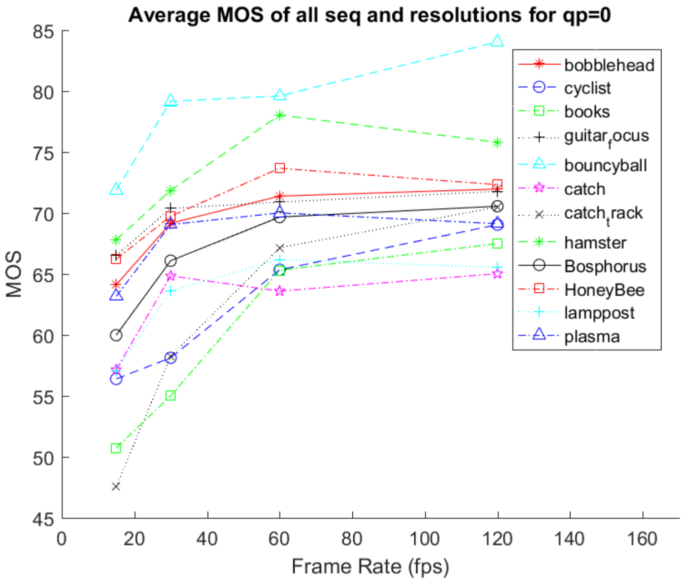


Figure 3.16: MOS versus frame rate for different contents in IVC-HFRVQA-II database with lossless compression(QP=0).

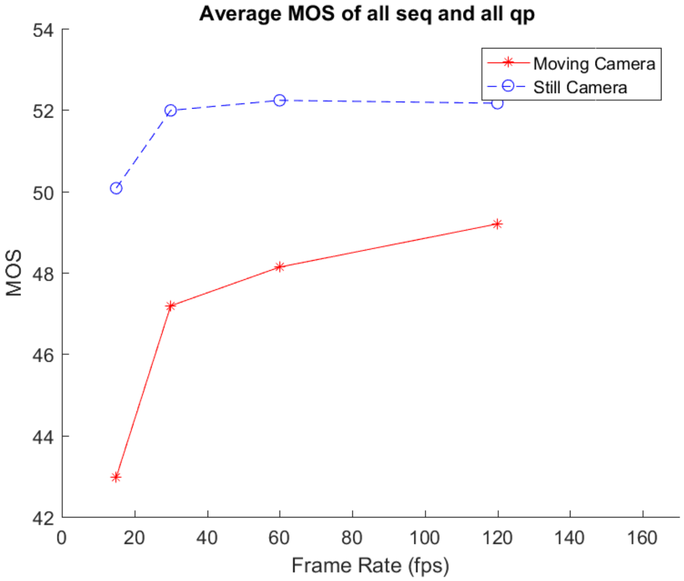


Figure 3.17: MOS versus frame rate for different videos in IVC-HFRVQA-II grouped by camera motion.

3.3 Test of VQA Models Blind to Frame Rate Variation

Before developing frame rate dependent VQA models, we first examine the performance of state-of-the-art VQA methods that do not consider the impact of frame rate directly. We evaluate the performance of well-known VQA methods including SSIM[21], MS-SSIM[23], and VQM[72].

For performance evaluation, we used the perceptual quality assessment results obtained from our subjective study on IVC-HFRVQA-II database as the ground truth. The quality scores for videos in the IVC-HFRVQA-II database is calculated by using three aforementioned methods and are compared against the MOS values. In our work, the MOS values, from the best to the worst, range from 100 to 0 while the score range of the objective models may be in different ranges. We scaled all scores from each method to be in the same range of 100 to 0.

SSIM: SSIM is a quality metric originally designed for IQA but has also been used for VQA by comparing frames in the reference and test videos one by one and averaging the per-frame scores as an overall quality of the video. The scatter plot of the SSIM scores versus MOS values for the videos in the IVC-HFRVQA-I database is displayed in Figure 3.18, where it can be seen that SSIM has low correlation with human opinions across different frame rates. This disparity between subjective and objective scores is because SSIM does not consider the impact of frame rates on perceptual quality.

As it can be seen in Figure 3.18, SSIM does not capture the quality degradation made by resolution change. This can be observed from the scatter plot of MOS values vs. SSIM scores for different resolutions. This can be explained by considering that SSIM is blind to frame rate and resolution changes and cannot capture the perceptually important differences caused by frame rate and resolution change.

VQM:VQM is a well-known VQA method based on the comparison of coefficients of encoded frames in the Discrete Cosine Transform (DCT) space. VQM considers local contrasts and contrast sensitivity. VQM produce a score greater or equal to zero, where zero denotes the best quality. The quality scores by using VQM are calculated for IVC-HFRVQA-II database and normalized in the range of 100 to 0 for the best and worst quality, respectively. Figure 3.19 shows the VQM scores vs MOS values, where VQM also fails to take into account of frame rate and resolution changes.

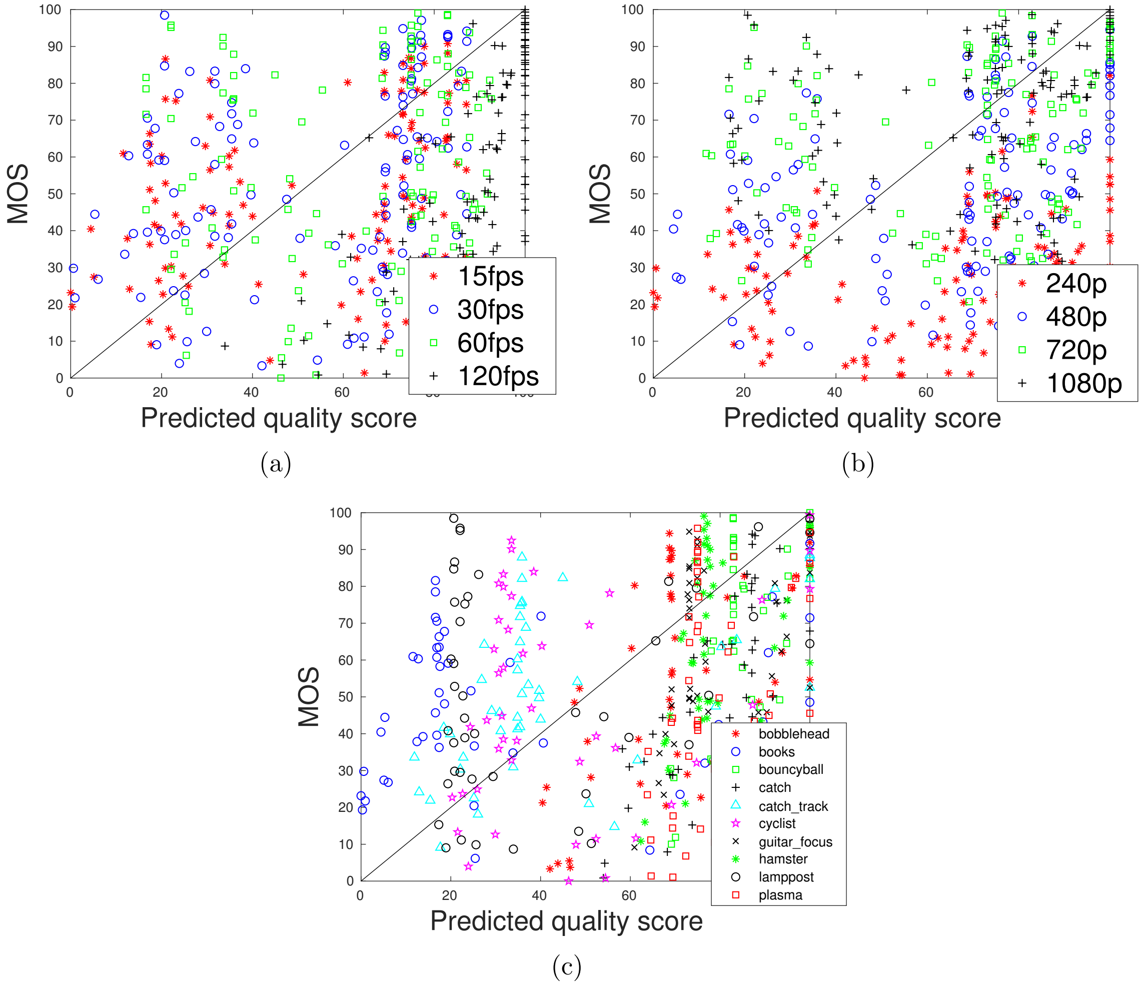
MS-SSIM: MS-SSIM is a variation of SSIM which showed a significant improvement over SSIM in predicting quality of still images and videos. Similar to SSIM, MS-SSIM is originally designed for IQA task, but has also been used for VQA by pooling per-frame scores. We calculated the MS-SSIM scores for videos in the IVC-HFRVQA-II database and normalized the scores to the range of 100 to 0 for the best to worst quality. Figure 3.20 shows the scatter plot of MS-SSIM scores vs MOS values, where MS-SSIM generally works better than SSIM and VQM in predicting quality, especially at low quality range. However, the overall quality prediction performance is still poor. Although MS-SSIM is a multi-scale approach, the cross-scale weighting is not adapted to the resolution of the video or the viewing conditions. In addition, it is also blind to frame rate changes.

Table 3.8: The performance of the well-known VQA methods on predicting the quality of videos in IVC-HFRVQA-II dataset.

Method	SRCC	PLCC	RMSE
SSIM	0.3298	0.2907	33.27
VQM	0.2089	0.2501	30.60
MS-SSIM	0.6423	0.6704	39.03

Table 3.8 shows the performance of the aformentioned VQA methods in predicting video quality for IVC-HFRVQA-II database. The results show that even the best-performing MS-SSIM method only has limited capability at predicting subjective quality evaluation.

The test results of state-of-the-art VQA models on the IVC-HFRVQA-II database suggest that existing VQA methods are very limited at producing meaningful quality prediction for video content cross different frame rates and resolutions, suggesting deeper investigation on the impact of frame rate and resolution changes is highly desirable.

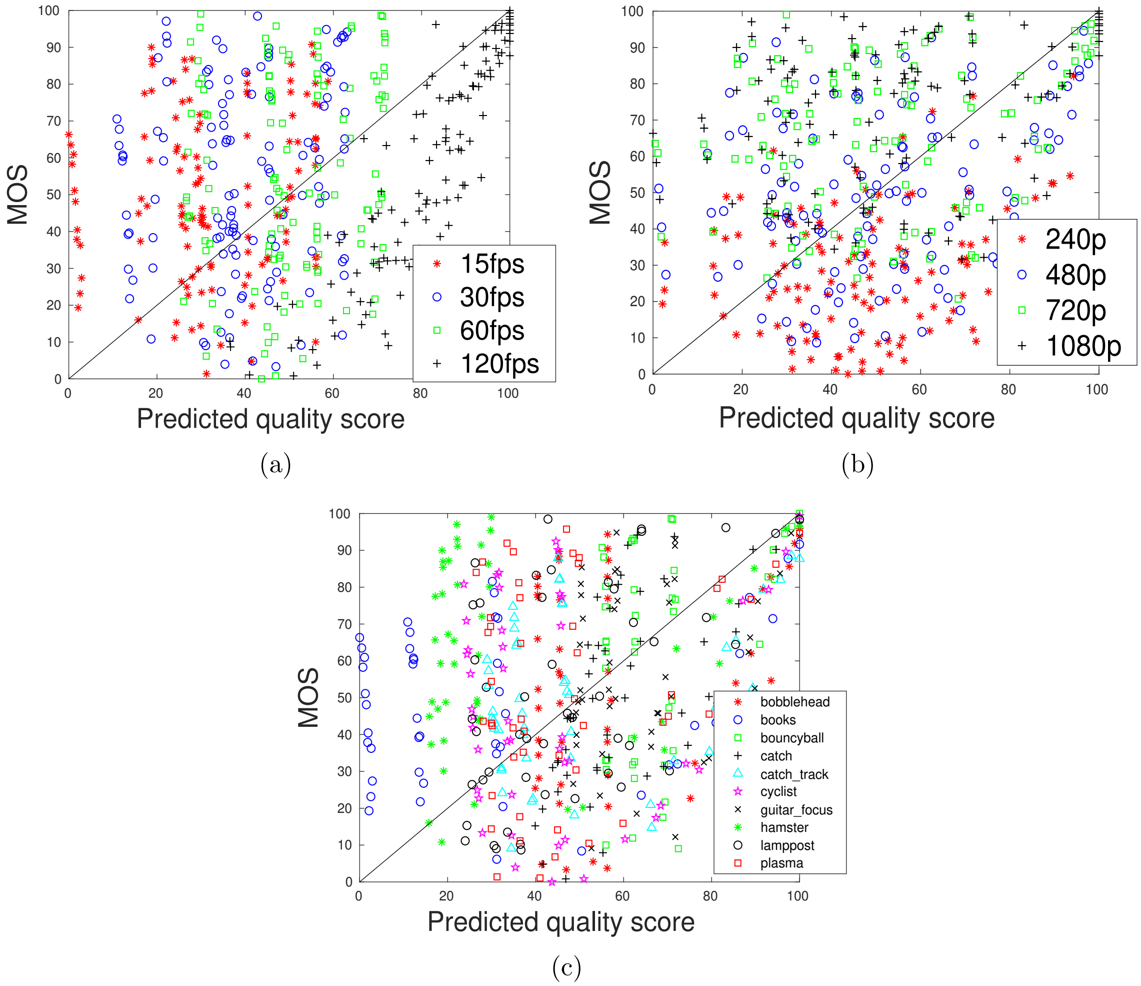


(a)

(b)

(c)

Figure 3.18: Scatter plot of MOS values versus SSIM scores for the IVC-HFRVQA database.

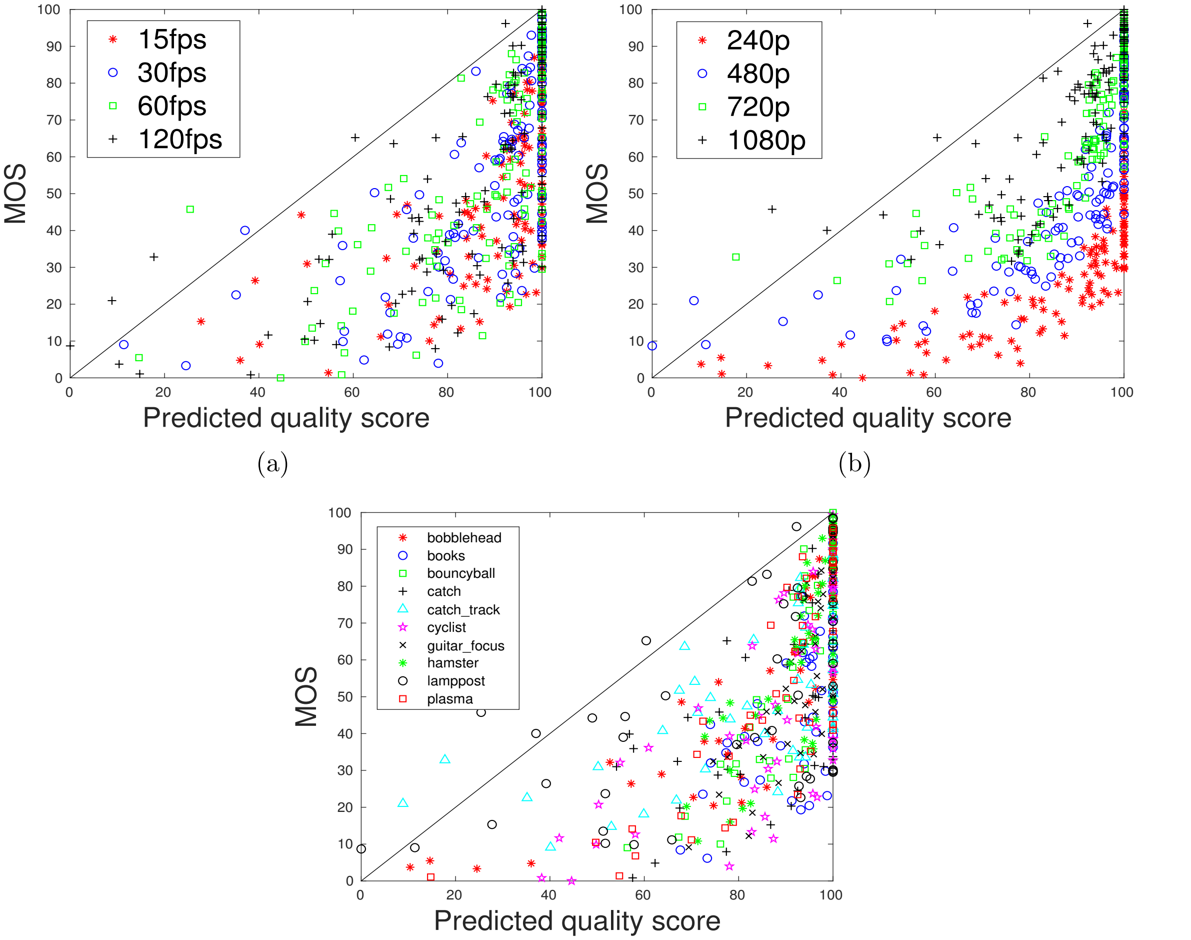


(a)

(b)

(c)

Figure 3.19: Scatter plot of MOS values versus VQM scores for the IVC-HFRVQA-II database.



(a)

(b)

(c)

Figure 3.20: Scatter plot of MOS values versus MS-SSIM scores for the IVC-HFRVQA-II database.

3.4 Summary

We constructed two video databases and subsequently performed subjective testing on the quality of videos at different quantization levels, spatial resolutions and frame rates. We have several interesting observations of the data which can be used for creating objective VQA model. In overall, quality has direct relation with frame rate, however the impact is more significant for frame rates lowever than 60 fps. Also, the impact of resolution change is more dominant to frame rate change specially for high frame rates.

We also evaluated some well-known VQA methods in predicting video quality in our databases. We showed that the performance of traditional VQA methods that do not take frame rate and resolution change into account is far from the human opinion obtained from subjective studies.

Chapter 4

Statistical Modeling of Motion Smoothness for Cross-Frame Rate Video Quality Assessment

4.1 Statistical Motion Smoothness Factor

Natural Scene Statistics (NSS) have attracted significant amount of attention in recent years [73] and have been used in a number of applications such as object classification in images [74], image coding [75], image representation [76], NSS have also been used in quality assessment of image and videos [77, 78, 79]. It has been long hypothesized that the biological visual systems are highly adapted to NSS during the evolution and development processes. As a result, NSS based analysis can help understand the behavior of biological processes in the human visual system on how visual information is extracted and encoded.

Motion smoothness is one of such statistical information from both the perspective of vision science and the engineering practice of video quality assessment. In [80] motion smoothness is used to analyze natural image sequences, where it has been shown that temporal distortions are related to the loss of local phase correlations. This characteristic is used for reduced reference video quality assessment in [81] and the results show that motion smoothness is a useful feature in estimating the quality of distorted videos. Here we opt to use motion smoothness to estimate motion distortions caused by frame rate changes in cross-frame rate VQA. We hypothesize that motion smoothness will decrease with reducing frame rate. Thus the comparison of motion smoothness between the reference and distorted videos may be used as an indicator of quality degradations caused by frame rate change.

For calculating motion smoothness, we first assume that there is a rigid motion for a one dimensional signal $f(x)$ and this motion can be modeled as

$$h(x,t) = f(x + u(t)) + b(t) \tag{4.1}$$

where $b(t)$ is the time varying background luminance which is approximately constant in a short period of time and $u(t)$ is the motion over time.

Consider a family of complex wavelets of the form $w(x) = g(x)e^{jw_c x}$, where $g(x)$ is varying slowly with x and w_c is the wavelet center frequency. The variations of a mother wavelet $w(x)$ can be generated as follows :

$$w_{s,p} = \frac{1}{\sqrt{s}}(\frac{x-p}{s}) = \frac{1}{\sqrt{s}}(\frac{x-p}{s})e^{jw_c x/s} \tag{4.2}$$

where s and p are scale and shift factors, respectively. Then the complex wavelet transform of the signal $f(x)$ can be computed as

$$\begin{aligned} F(s,p) &= \int_{-\infty}^{\infty} f(x) w_{s,p}^{*}(x) dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} F(w) \sqrt{s} G(sw - w_c) e^{jwp} dw, \end{aligned} \tag{4.3}$$

where $F(w)$ and $G(w)$ are Fourier transforms of $f(x)$ and $g(x)$, respectively. Applying this transform on the motion model of Eq. 4.1 leads to

$$\begin{aligned} H(s,p,t) &= \int_{-\infty}^{\infty} F(w) \sqrt{s} G(sw - w_c) e^{jw(p+u(t))} dw \\ &\approx F(s,p) e^{j(w_c/s)u(t)} \end{aligned} \tag{4.4}$$

Take a logarithm on both sides, we have

$$\log H(s,p,t) \approx \log F(s,p) + j(w_c/s)u(t) \tag{4.5}$$

The imaginary part of the above equation has a linear relationship with respect to motion $u(t)$. To relate motion smoothness with this complex wavelet transform, we examine complex wavelet coefficients starting from a time instance t_0 and sample the sequence at consecutive time steps $t_0 + n\Delta t$ for n=0,1,...,N. The N-th order temporal correlation function is defined as

$$L_N(s,p) \approx \sum_{n=0}^{N} (-1)^{n+N} \left({N \atop n} \right) \log H(s,p,t_0 + n\Delta t) \tag{4.6}$$

Using (4.5), it can be shown that when the motion is (N-1)-th order smooth (meaning that all derivatives of $u(t)$ in degree higher than N are zero), $L_N(s,p) \approx 0$. However this result is for the ideal case of perfectly smooth motion. Real-world videos deviate from the ideal case, and such deviation may be used as a measure of motion smoothness.

Meanwhile we define a temporal energy function as

$$M_N(s,p) \approx \sum_{n=0}^{N} \left({N \atop n} \right) \log H(s,p,t_0+n\Delta t) \tag{4.7}$$

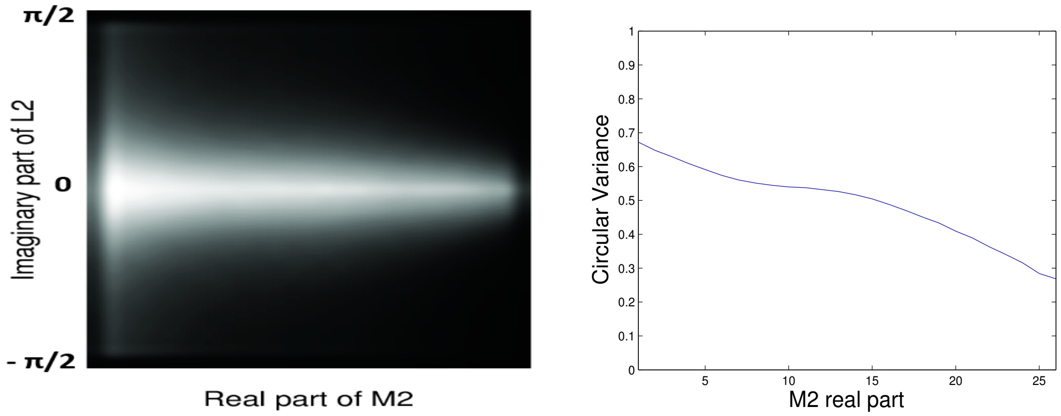
By examining the temporal correlation function (L_N) and temporal energy function (M_N) jointly, one can observe how temporal motion smoothness varies as a function of local signal energy.

An example of the joint histogram is shown in Figure 4.1(a), where brighter indicates more frequent occurrence. To observe the trend of motion smoothness with respect to the local signal energy, the Circular Variance(Circular Variance (CV)) [82, 83] of the two dimensional histogram is calculated for each column, which results in a measure of variation in the histogram for each energy level (column). Given the 2D joint histogram, the CV is calculated for each column of the histogram by

$$CV_q = 1 - \frac{|\sum_{p=1}^{M} h_{p,q} e^{j\theta_p}|}{\sum_{p=1}^{M} h_{p,q}}, \tag{4.8}$$

where θ_p and $h_{p,q}$ are the center angle and height of bin p in column q and M is the number of bins in the histogram.

The trend of the CV versus the energy level is shown in Figure 4.1(b). In the case of perfect motion smoothness, the phase prediction error or the imaginary part of L2 should be zero. As a result, the whole histogram is concentrated at the center zero line regardless of the energy measure. Diffusion from the central line indicates reduction in motion smoothness, and the level of diffusion is measured by the CV curve at different energy level, as examplified by Figure 4.1(b).



(a) 2D histogram

(b) CV

Figure 4.1: Temporal motion smoothness by (a) joint histogram of (L_N, M_N); and (b) measure of circular variance on columns of joint histogram.

The normalized area under CV curve quantify the overall temporal motion smoothness (TMS) as

$$TMS = \frac{\sum_{q=1}^{K}(1 - CV_q)}{K}, \tag{4.9}$$

where K is the number of columns.

4.2 Evaluation of Motion Smoothness Factor on IVC-HFRVQA-I database

The proposed measure of motion smoothness is applied to the IVC-HFRVQA-I database to estimate motion smoothness. Figure 4.2 shows the histogram of the imaginary part versus energy for all seven contents of the IVC-HFRVQA-I database in addition to the CV plots for each sequence.

To observe the trend of motion smoothness in different frame rate, Figure **??** shows the joint histogram along with the CV for each content and for three frame rates of 15, 30, and 60 fps.

It appears that increasing the frame rate leads to less variance in motion and smoother motion. As we increase the frame rate from 15 to 30 and then to 60fps (from column 2 to 4), the histogram concentrates on the middle line of zero phase prediction error representing smooth motion. From the CV curves in the last column, it can be observed that the smoothness decreases as we decrease the frame rate, however the trend varies for different contents. For example the difference in sequence 1 and 3 are more significant than the others. Interestingly, our subjective test shows these two sequences produce more MOS improvement by increasing frame rate. Another observation is that for sequence 5 (talk) the motion is low, and the smoothness of motion is not varying significantly across different frame rates.

To investigate the relationship between temporal motion smoothness and the compression level controlled by QP values, different levels of compression is analyzed for three different frame rates. Figures 4.5 to 4.11 show the trend for CV and for different QP levels and different frame rates. It can be seen that motion smoothness is almost the same for different QP levels. This suggests that motion smoothness may not be an ideal measure to capture the severeness of compression artifacts.

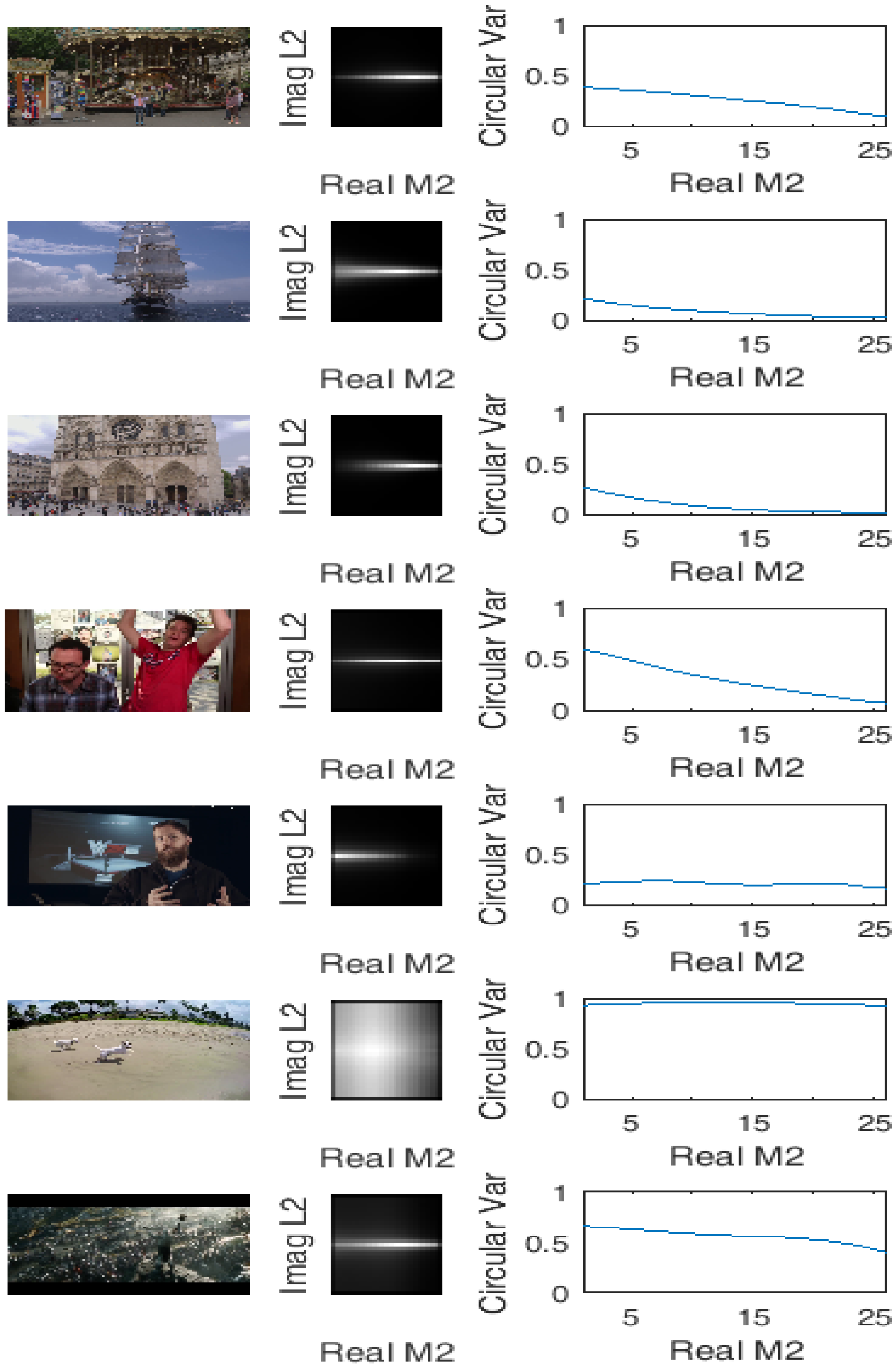


Figure 4.2: Temporal motion smoothness by 2D joint histogram and measure of circular variance for different content at 60fps and 480p resolution. First column: snapshot of video; Second column: 2D joint histogram; Third column: CV curve.

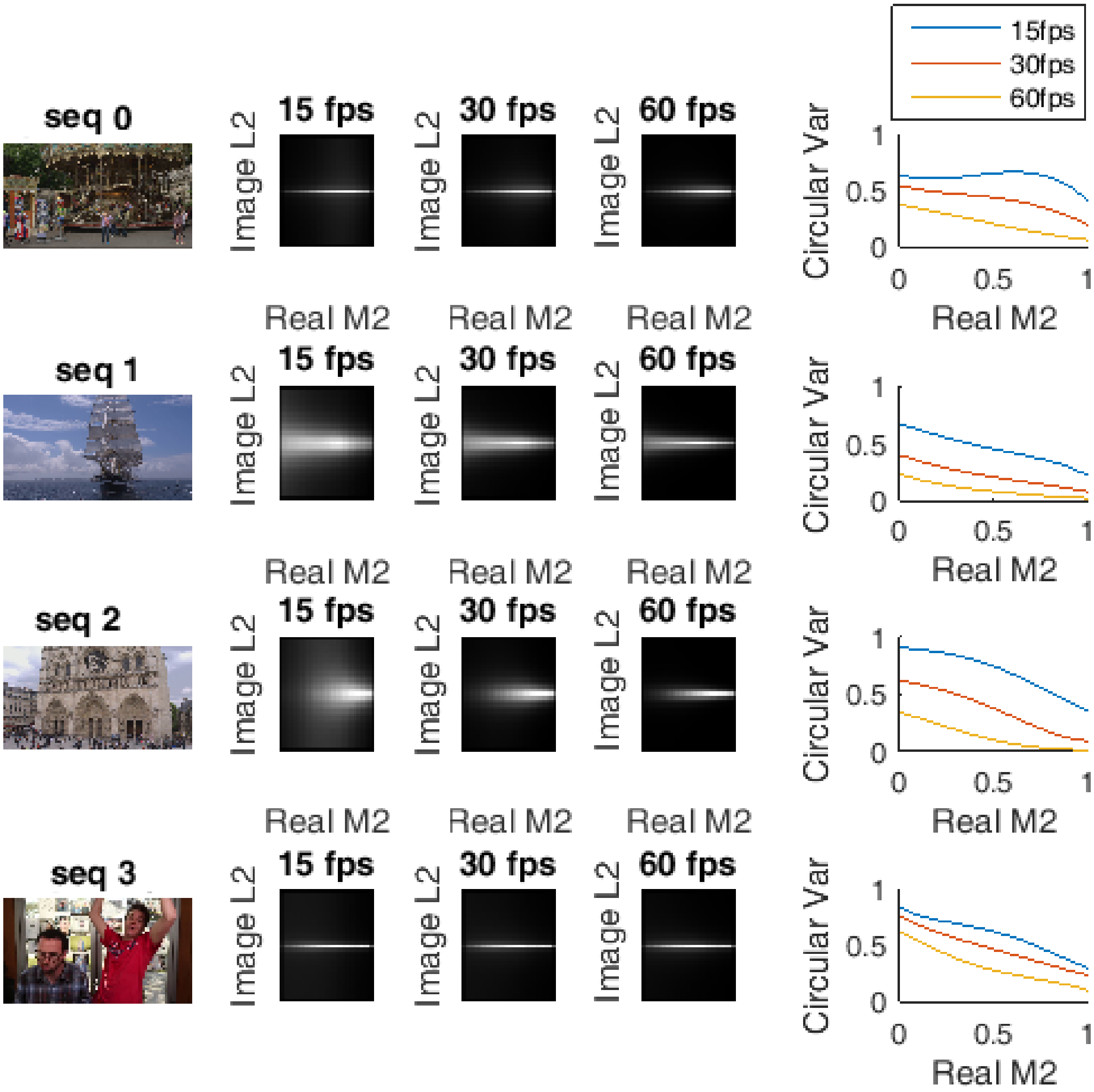


Figure 4.3: Temporal motion smoothness by 2D joint histogram and measure of circular variance for different frame rates.

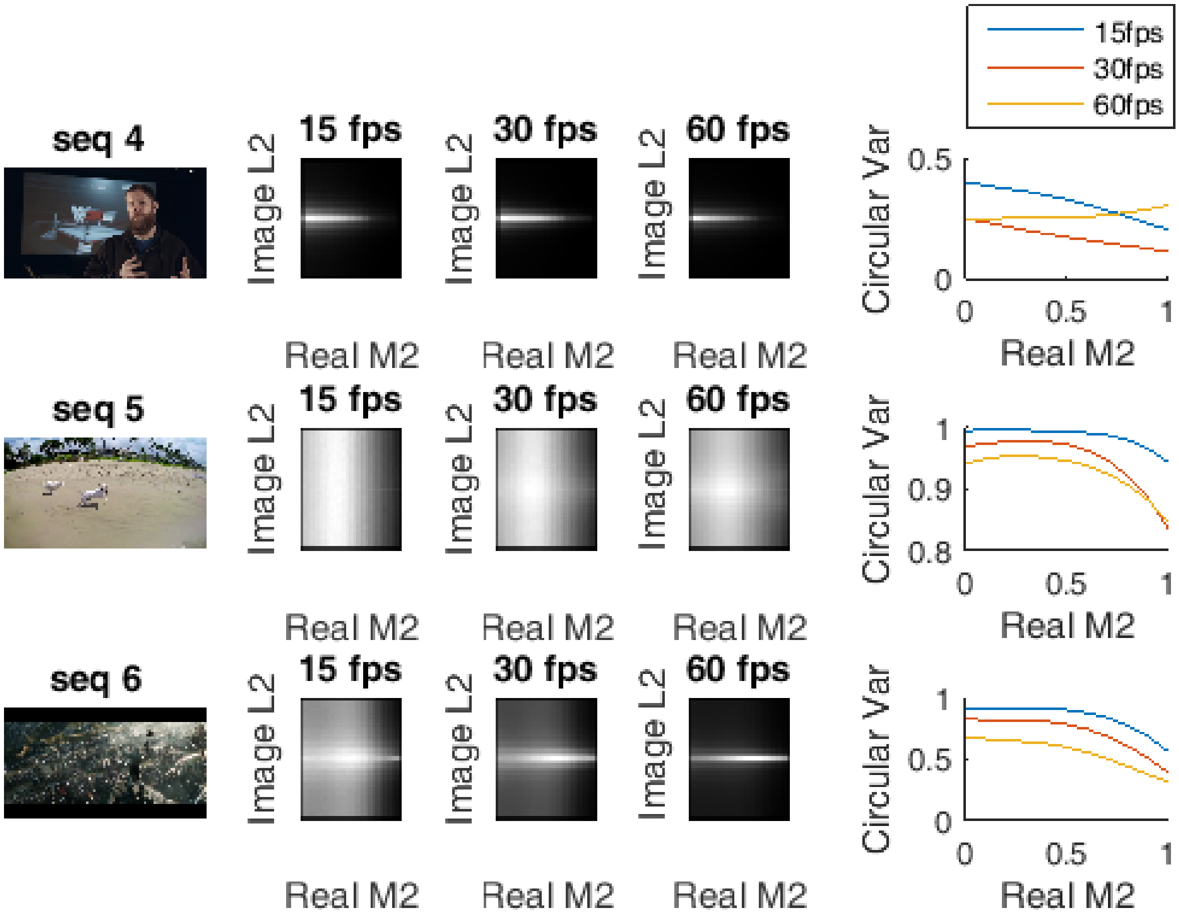


Figure 4.4: Temporal motion smoothness by 2D joint histogram and measure of circular variance for different frame rates(cont'd).

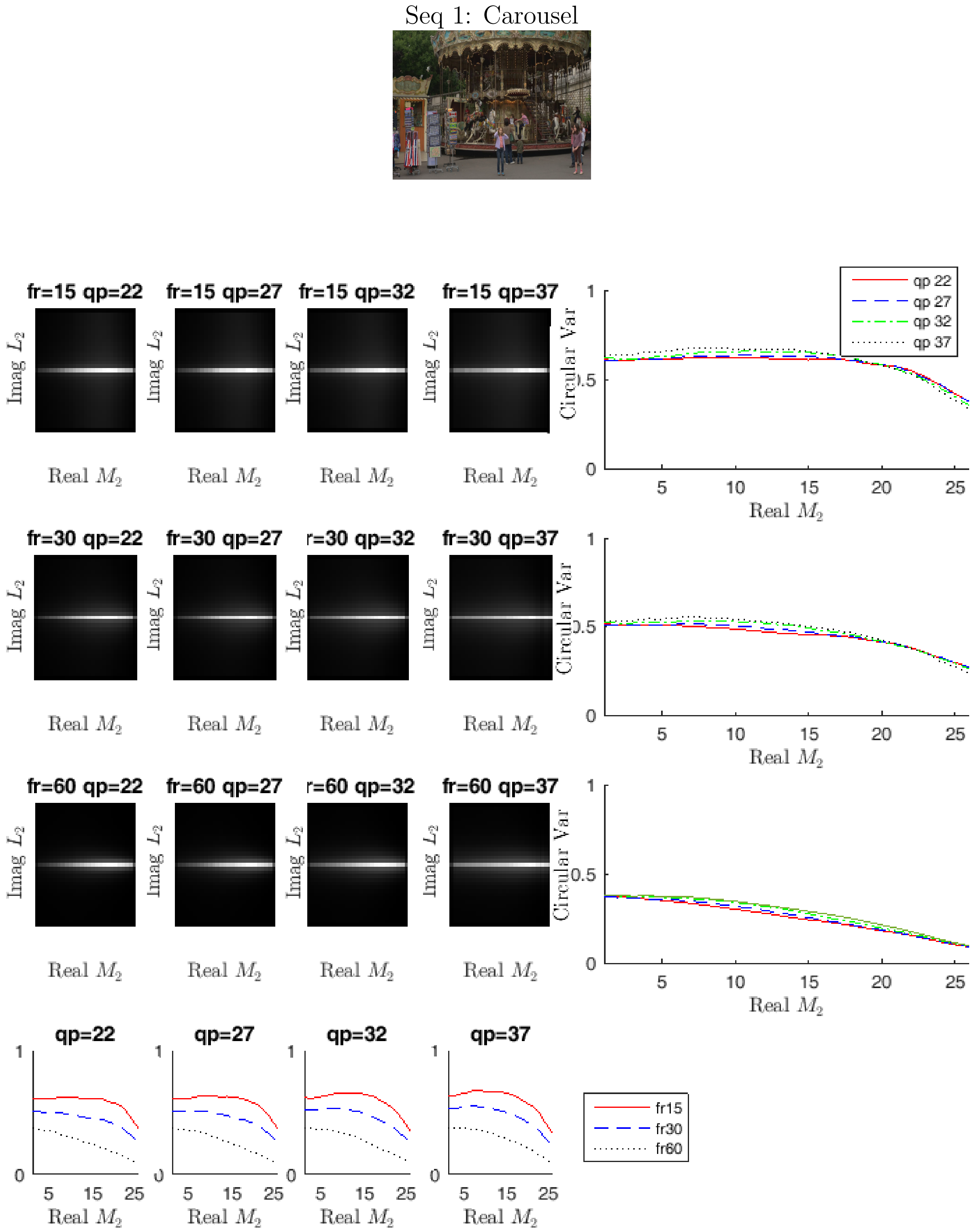


Figure 4.5: Temporal motion smoothness by 2D joint histogram and circular variance for Sequence 1.

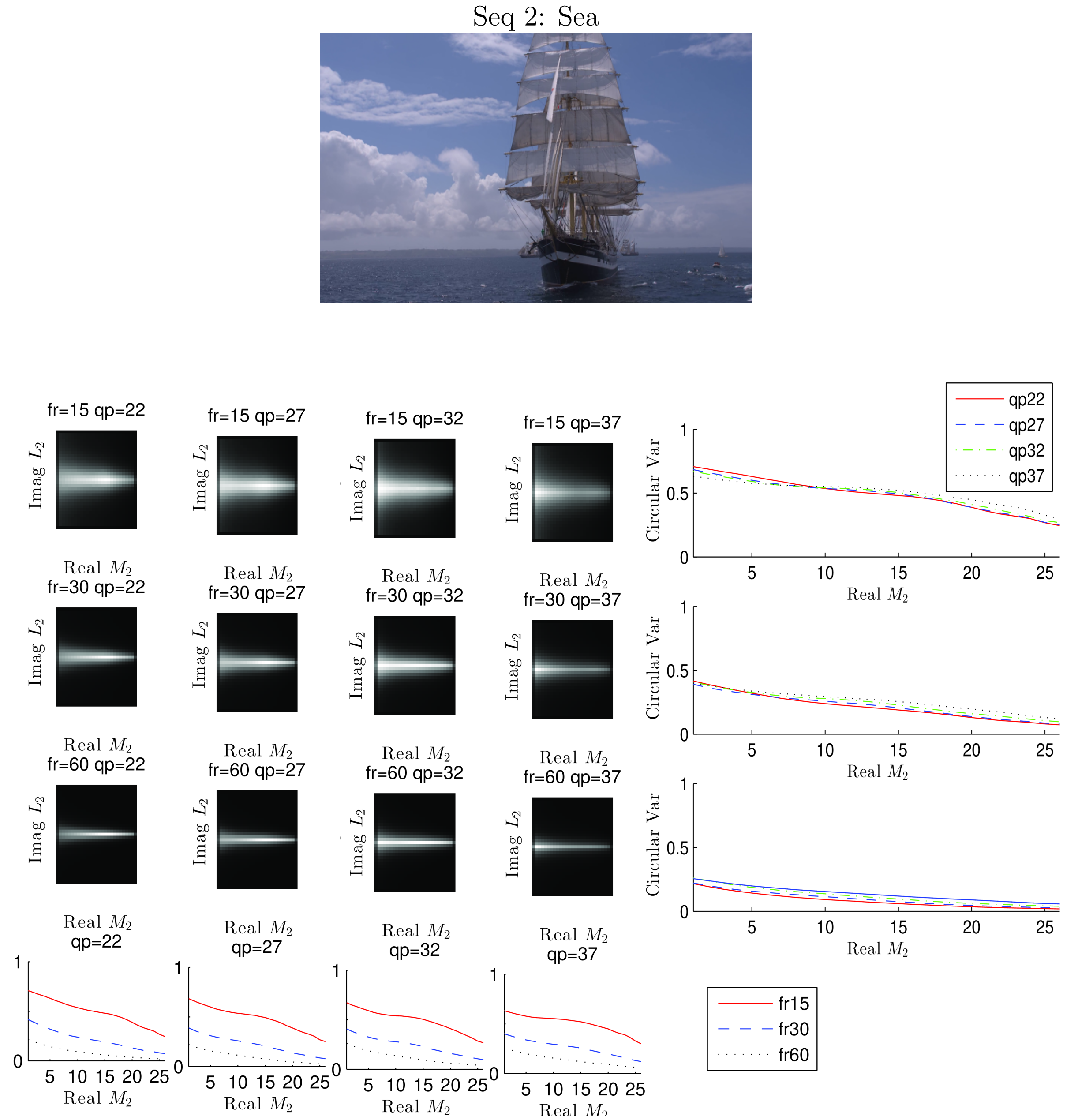


Figure 4.6: Temporal motion smoothness by 2D joint histogram and circular variance for Sequence 2.

Seq 3: Notre Dame

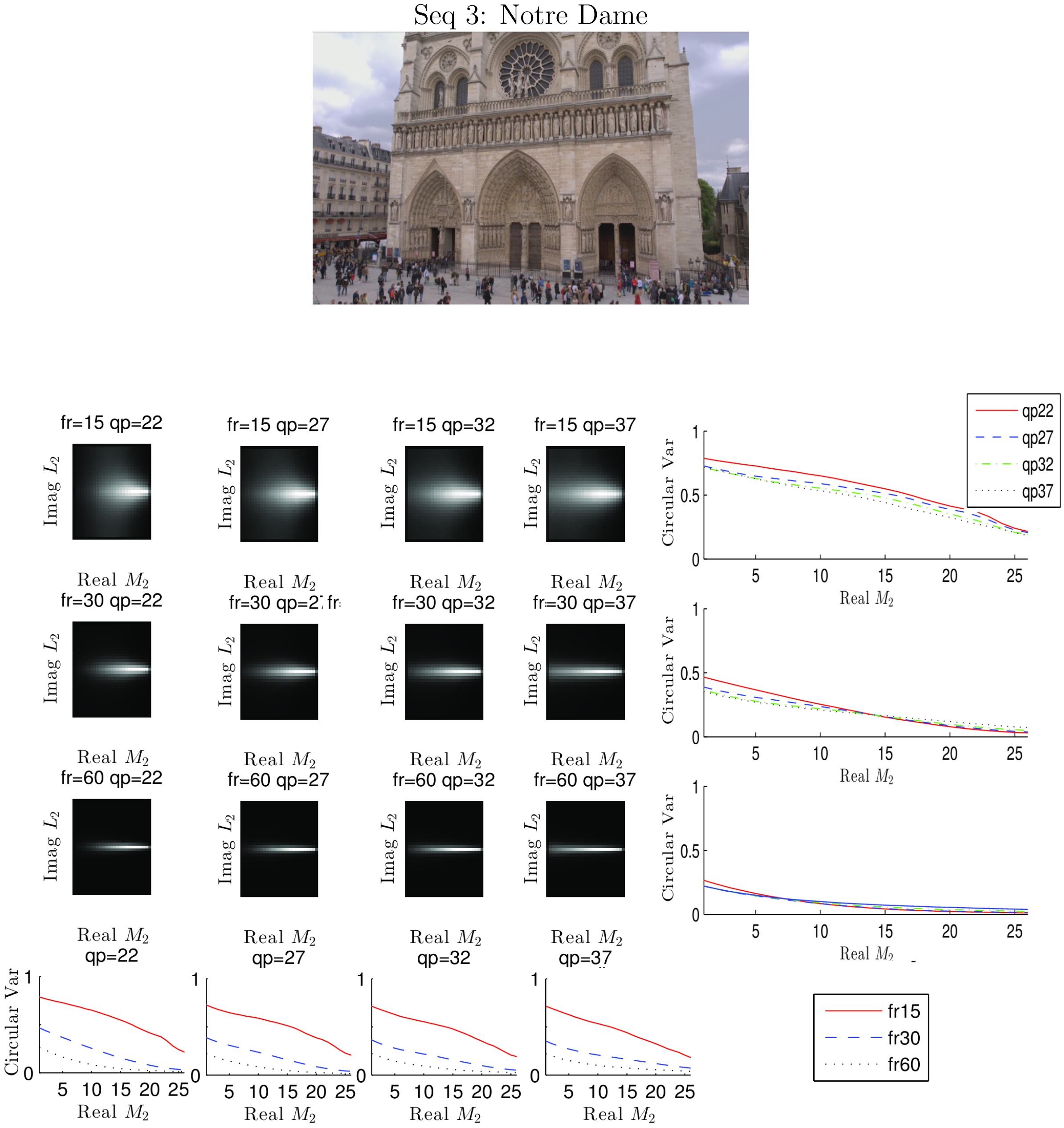


Figure 4.7: Temporal motion smoothness by 2D joint histogram and circular variance for Sequence 3.

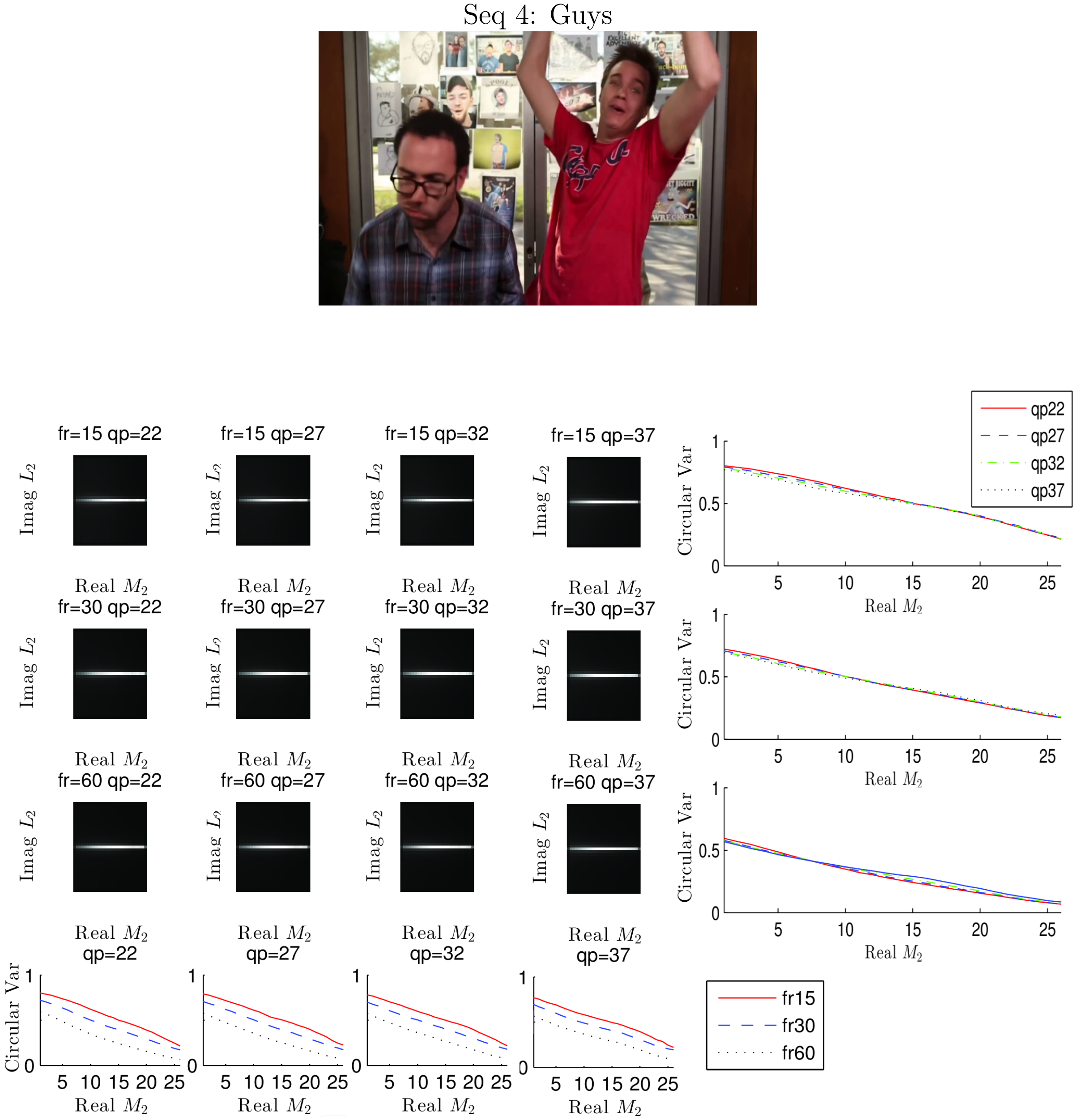

Seq 4: Guys

Figure 4.8: Temporal motion smoothness 2D histogram and circular variance for Sequence 4.

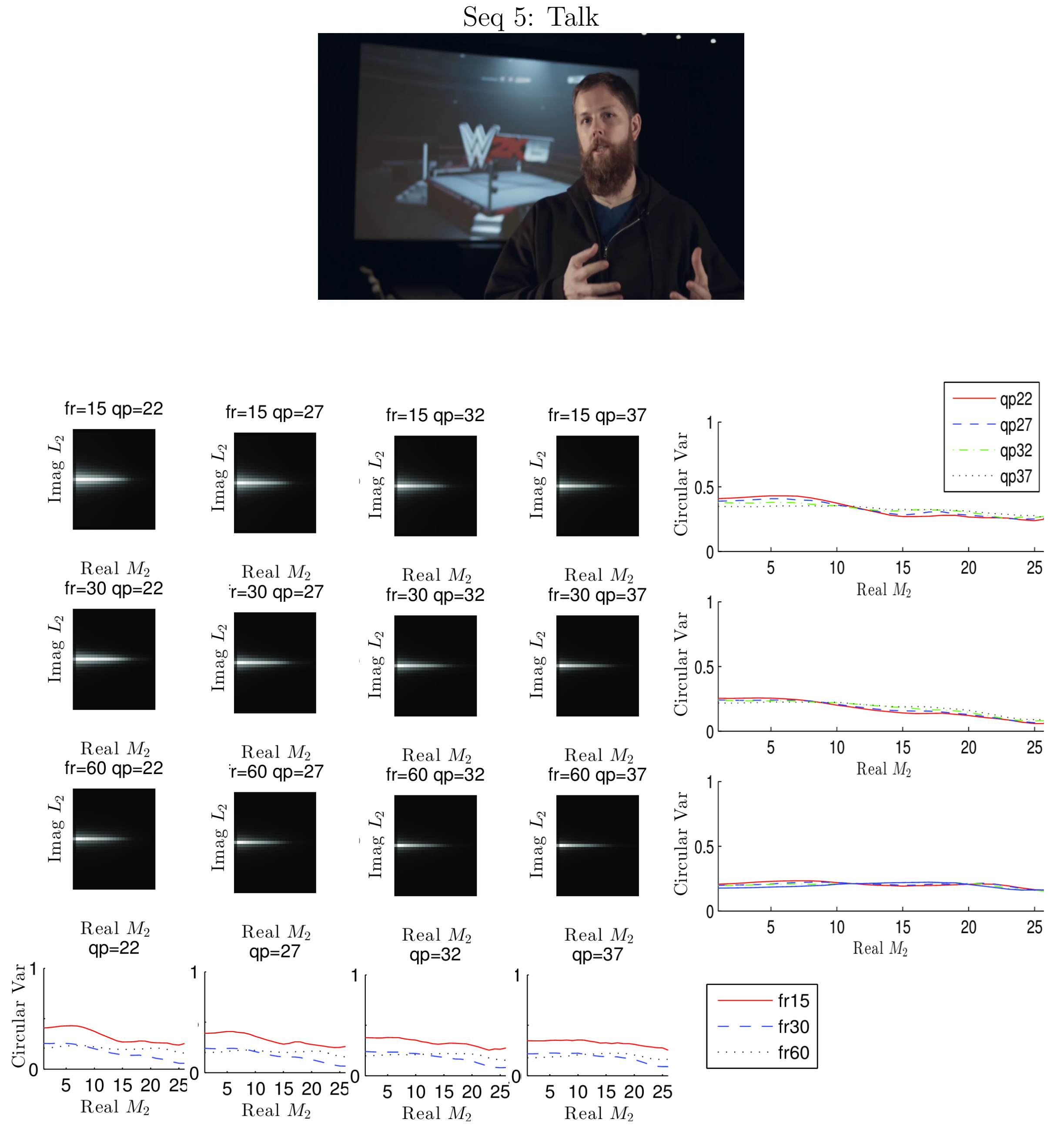


Figure 4.9: Temporal motion smoothness by 2D joint histogram and circular variance for Sequence 5.

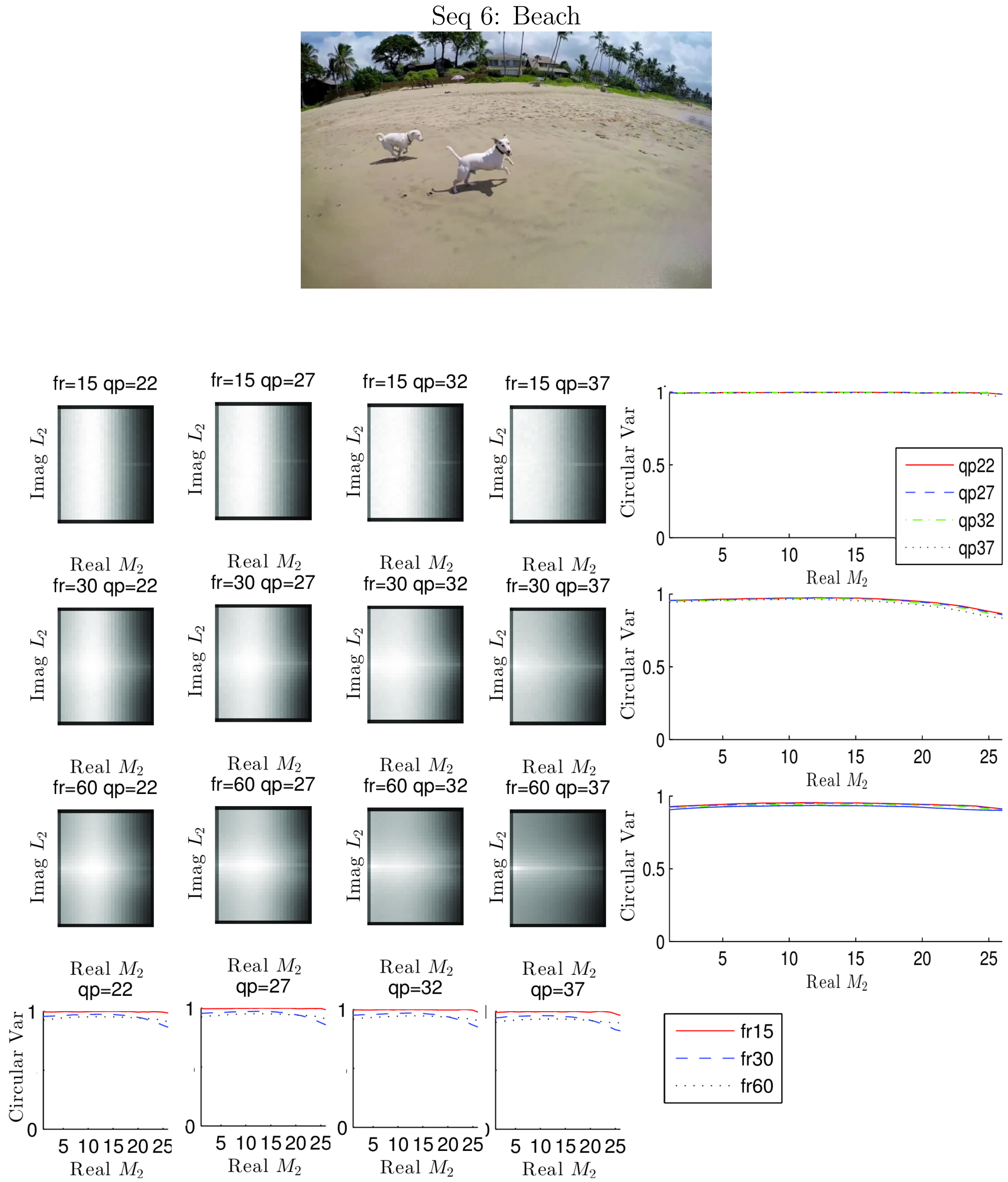


Figure 4.10: Temporal motion smoothness by 2D joint histogram and circular variance for Sequence 6.

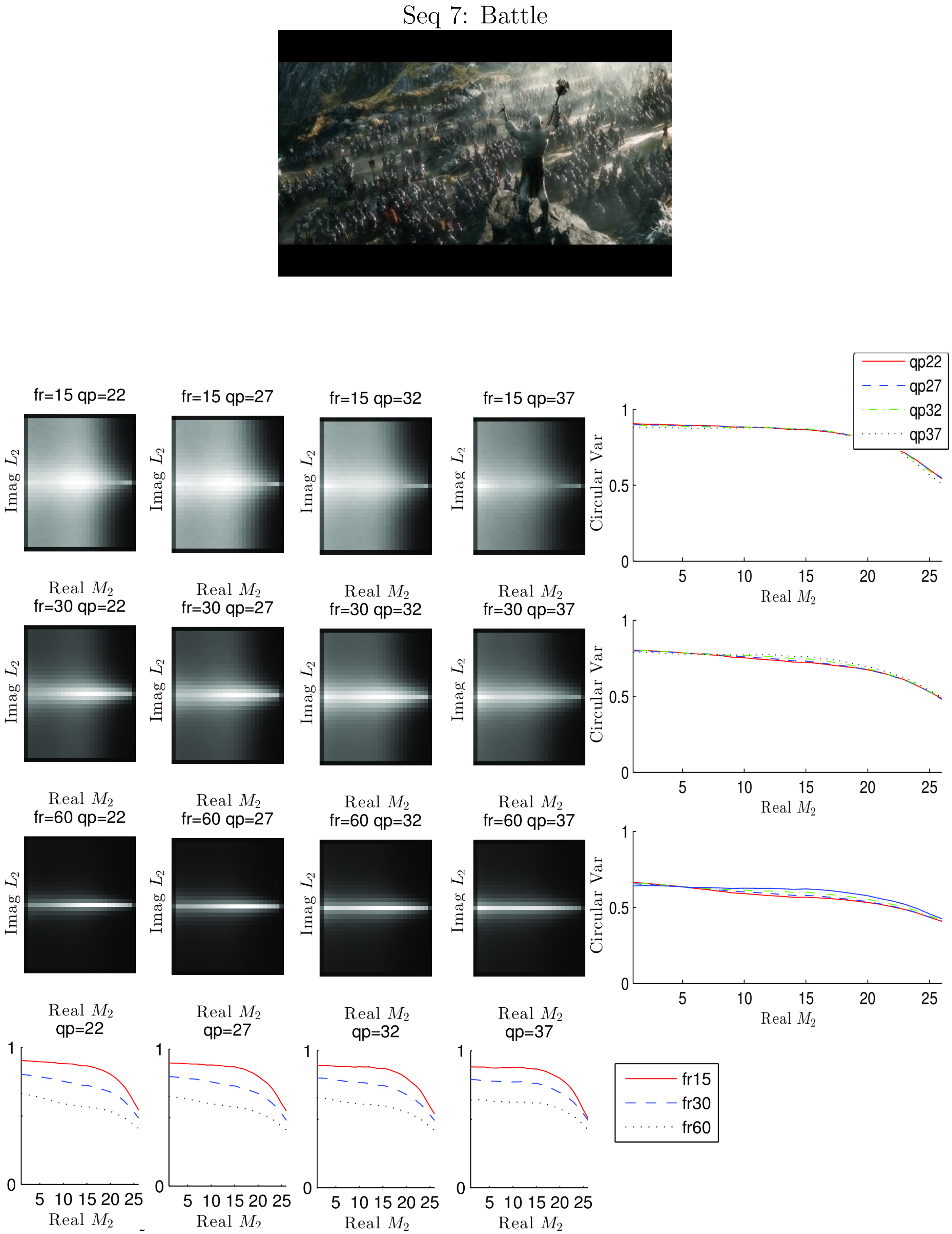


Figure 4.11: Temporal motion smoothness by 2D joint histogram and circular variance for Sequence 7.

4.3 Evaluation of Motion Smoothness Factor on IVC-HFRVQA-II database

We evaluate the proposed TMS measure on IVC-HFRVQA-II database. In this database, each source sequence is originally at 120 fps and is converted to lower frame rate test sequences at 60 fps, 30 fps, and 15 fps, respectively. We focus on the frame rate change between lossless compressed videos (qp=0) only in IVC-HFRVQA-II database. Using the database, we first examine how the proposed TMS measure correlates with video frame rate and human subjective QoE for individual video content. We then investigate further on motion-based content dependencies.

4.3.1 Validation

To better understand and to demonstrate the proposed motion smoothness measure, we examine how the joint histogram and its corresponding CV change with respect to different frame rates for videos in IVC-HFRVQA-II in Figs. 4.12, 4.13, 4.14, and 4.15. It can be observed from Fig. 4.14, and 4.15 that regardless of the content variation, the effect of frame rate reductions is well captured by the departure of the CV curves of the distorted videos from the reference CV curves. Specifically, the CV curve generally moves away from the reference CV curve with the decrease in frame rate. This is further confirmed by the high Spearman rank-order correlation coefficient (SRCC) between the TMS factor and MOS shown in Table 4.1. The only exception appears to be the “hamster” sequence, where the proposed TMS factor is unable to distinguish the reference and distorted videos. The possible reason could be that the spatial variation in motion pattern and speed are very high, or the local motion pattern in high speed refresh rate may be too complicated to be fully captured by the phase correlation between complex wavelet coefficients.

Table 4.1: Correlation of TMS factor with frame rate and DMOS for individual video sequences of different motion types (global vs local motion) and spatial motion variation levels.

video sequence	motion type	spatial motion variation	SRCC of TMS vs. DMOS
bobblehead	global	low	-1
books	global	low	-1
bouncyball	local	medium	-1
cath_track	global	low	-1
cath	local	high	-0.8
cyclist	global	low	-1
guitar_focus	local	medium	-0.8
hamster	local	high	-0.4
lamppost	local	medium	-1
plasma	local	medium	-1
mean/std	-	-	-0.90 /0.18

4.3.2 Motion Content Dependency

Although the proposed TMS factor exhibits a high correlation with perceptual quality within each content, its behavior varies significantly across different videos as is evident in Fig. **??**. For example, for high motion videos such as the "cyclist", there is much larger variation from the ideal smooth motion. This motivates us to study motion-based content dependency of the proposed motion smoothness measure.

An important aspect of motion in the video is the presence of camera geometric transformation in the video acquisition process. Such camera motion transformations result in global motion in video. The global motion has important impact on the visibility of distortion in video and general perception of video quality [39]. For example, the blurring artifact is less visible in videos with globally very fast motion and such effects have been considered in existing video quality models [39].

We classify videos into two groups based on the presence of global motion, and computed

SRCC between the TMS factor and DMOS on each group. The experimental results are reported in Table 4.2. It can be seen that the proposed metric better predicts human opinions for videos containing global motion. This could be because the motion is more easily perceived in global motion videos as most regions of the frames are moving in a consistent manner. For the videos free of global motion (e.g. captured by static cameras), as moving regions are part of the frames only, the global statistics based TMS factor is less effective at reflecting the impact of such local changes in the overall perceptual quality.

Motion perception provides another important perspective that is missing in the proposed motion smoothness measure to study cross-frame-rate video quality assessment. Specifically, it has been shown that the perceptual motion information content is proportional to the strength of the relative motion and the inverse of global background motion [41]. A simple model to account for this relationship is given by

$$V = \frac{\sigma(\tilde{d})}{\mu(\tilde{d})} \tag{4.10}$$

where V represents the spatial motion variation, σ is the variation of frame difference, μ is the average of frame difference in pixels, and $\tilde{d}$ is the temporal frame difference. Intuitively, V increases as the motion statistics becomes more complex, and decreases as the uncertainty of motion perception μ increases. It is considered a measure of spatial motion variation, or perceptual motion information content (following the principle used in [41, 40]).

We use V to classify the videos used in this study into three classes-low, medium, and high spatial motion variation, as shown in Table 4.1. By conducting correlation analysis as reported in Table 4.2, we observe that the proposed metric works better for the videos with lower variation of motion across space. For the medium variation of motion videos,

Table 4.2: SRCC between $\tilde{S}$ and DMOS for different motion-based content types.

video group	**SRCC**
all	0.78
local motion	0.62
global motion	0.93
high spatial motion variation	0.71
medium spatial motion variation	0.87
low spatial motion variation	0.93

the correlation is close to low V class, and the accuracy of prediction drops significantly for the class of videos with high spatial variation in motion. This could be because for these videos, the proposed method calculates the average correlation of wavelet coefficients over the entire frame, while the motion is local and the human attention could be attracted to certain moving parts of the video frames. This suggests that segmenting the videos into different regions based on their motion characteristics and apply local TMS analysis is a direction worth deeper investigation in the future.

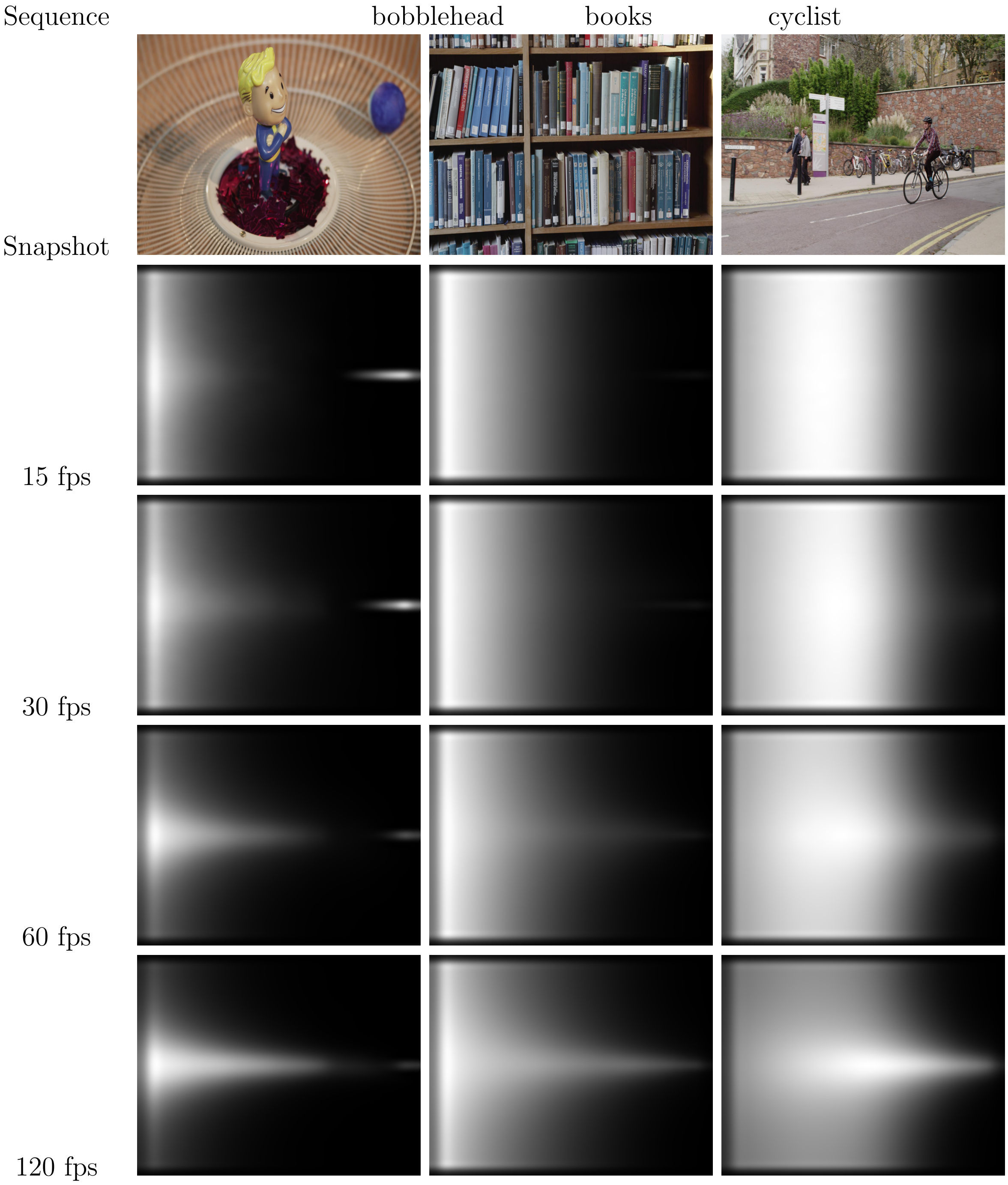


Figure 4.12: Temporal motion smoothness by 2D joint histogram of ($Re\{M_2\}$, $Im\{L_2\}$) for selected videos from BVI-HFR at four different frame rates [2]

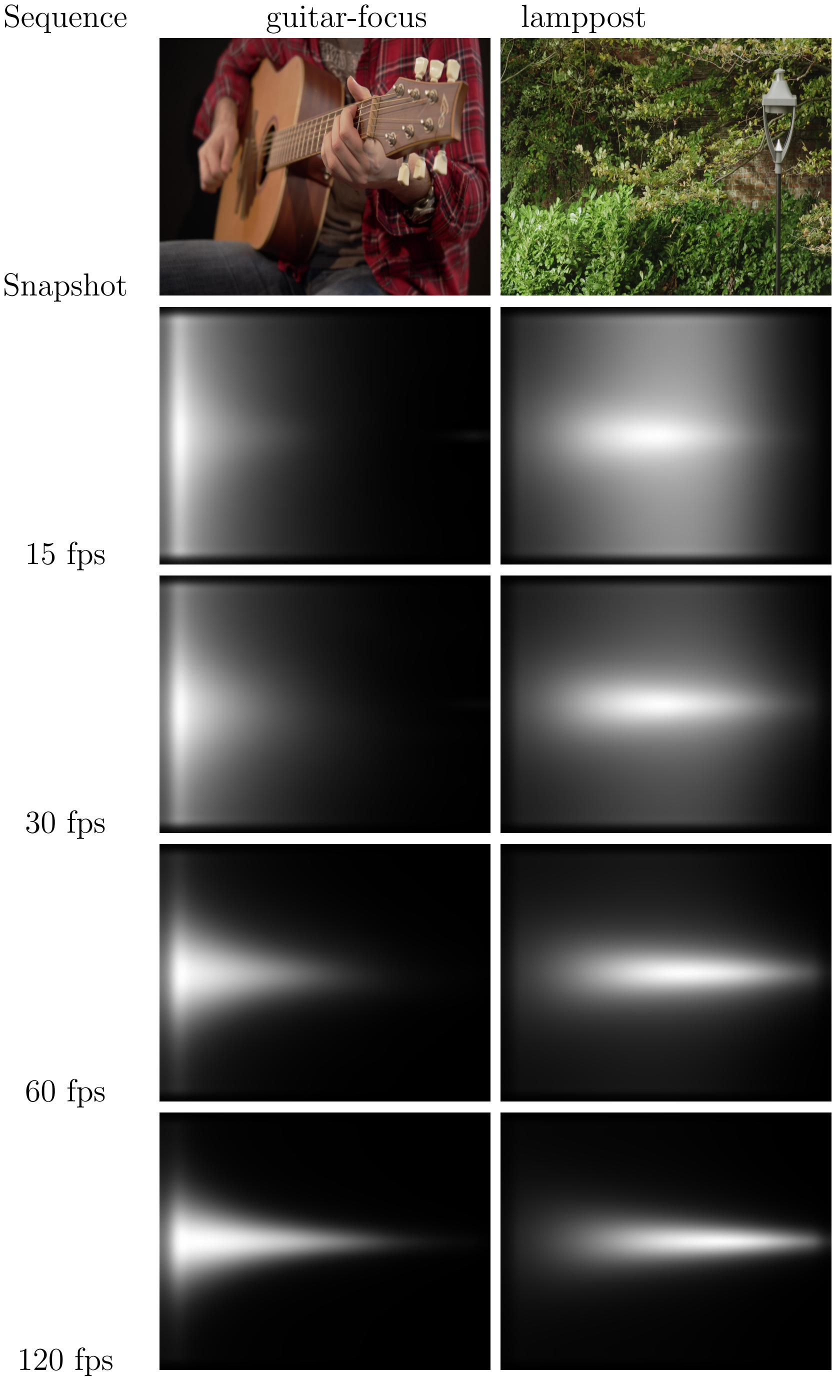


Figure 4.13: Temporal motion smoothness by 2D joint histogram of ($Re\{M_2\}$, $Im\{L_2\}$) for selected videos from BVI-HFR at four different frame rates [2]

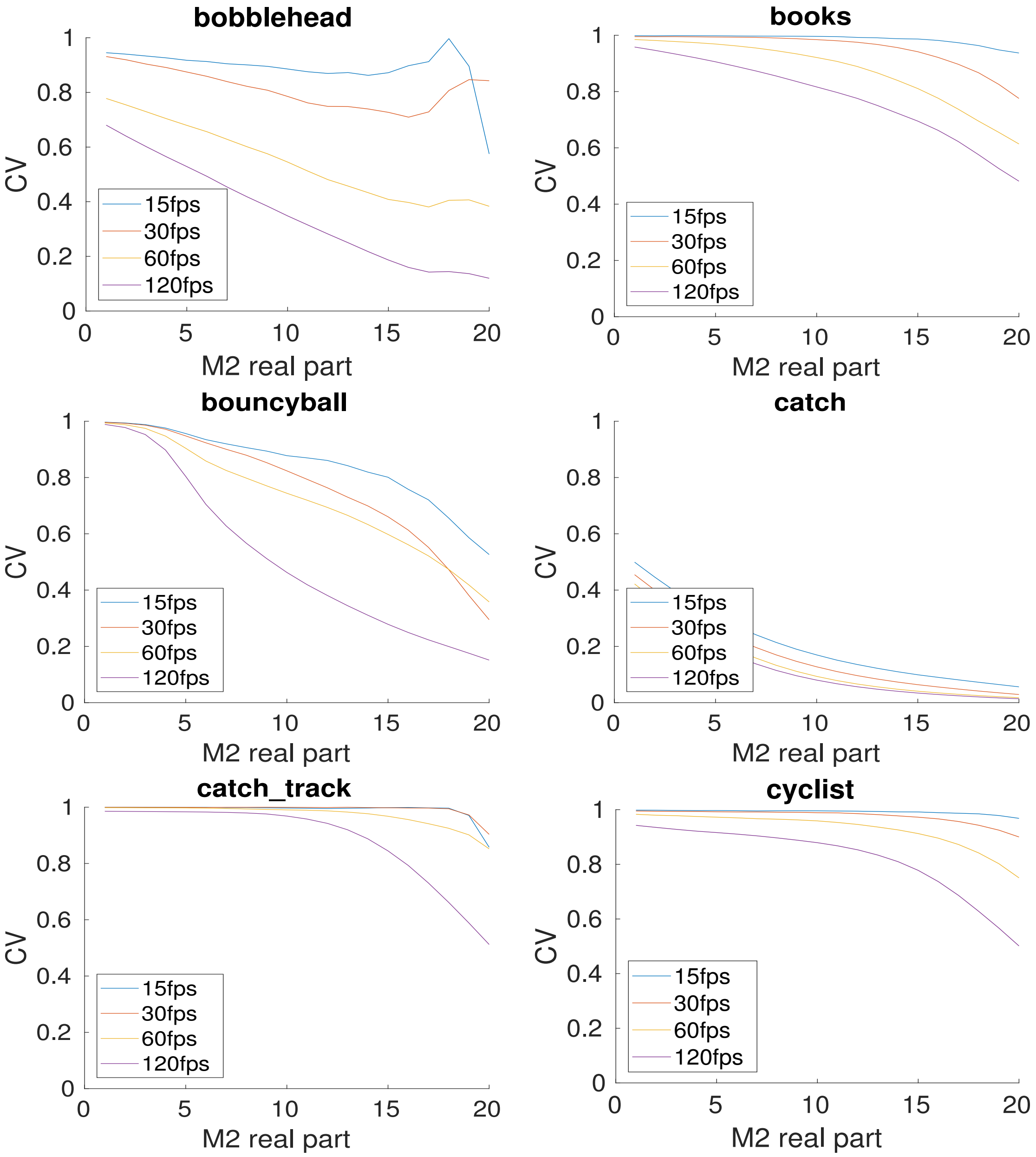


Figure 4.14: Circular variance curves of 2D joint histograms of $(Re\{M_2\}, Im\{L_2\})$ for the selected videos at different frame rates.

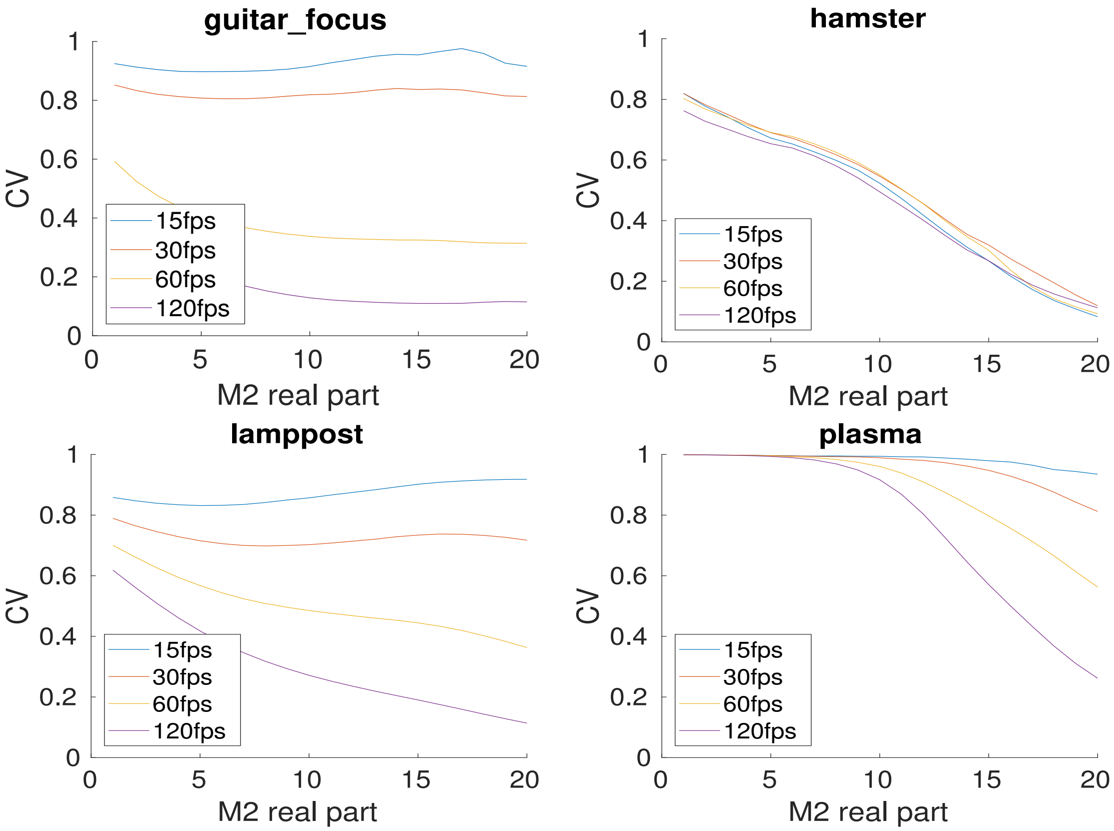


Figure 4.15: Circular variance curves of 2D joint histograms of $(Re\{M_2\}, Im\{L_2\})$ for the selected videos at different frame rates.

4.4 Summary

In this chapter, we investigated the impact of frame rate changes from the perspective of NSS and statistical analysis of motion in video. We employed a statistical model on the temporal correlations of complex wavelet transform coefficients to measure temporal motion smoothness of videos. We found that temporal motion smoothness monotonically decreases with the reduction of frame rate, which suggests that temporal motion smoothness may be a useful factor for VQA.

The statistical model investigated in this chapter revealed promising factors that may contribute to cross-frame rate VQA. However, more complete VQA models are to be developed that combine these models with other distortion measures to further improve the performance.

Chapter 5

Perceptual Aliasing Analysis for Cross-Frame Rate Video Quality Assessment

When considering a video as a three-dimensional signal with one temporal dimension and two spatial dimensions, frame rate is essentially the sampling rate in the temporal direction. Aliasing is a fundamental cause of signal degradation when the sampling rate is inadequate. In this chapter, we investigate various aliasing factors during frame rate changes and their relationship with perceptual video quality degradation. As both resolution change and frame rate change are common transforms to reduce the data rate of video, we explore the impact of frame rate in combination with variations in resolution changes.

5.1 Perceptual Aliasing Analysis

5.1.1 1D Temporal Aliasing Factor

Changing the frame rate of a video is equivalent to changing the sampling rate of the video in the temporal dimension. Therefore, distortions introduced by reducing the frame rate could be explored by analyzing the information loss caused by sampling rate reduction.

The most common method of frame rate reduction is performed by dropping a number of frames from the sequence of frames in the video. Frames are typically dropped in a uniform pattern depending on the original and target frame rates. An essential consideration in dropping frames or any down-sampling process is to consider the Nyquist theory of sampling, in which the sampling frequency (f_s) should be at least twice of the highest frequency component of the signal; otherwise aliasing occurs. Since images and videos usually contain a variety of frequency components in their frequency spectrum, obeying the Nyquist criteria in order to avoid aliasing requires the signal to be filtered by a low-pass anti-aliasing filter before down-sampling.

In practice, however, low-pass filtering is not commonly performed before frame rate reduction due to the added computational cost and memory requirement. This leads to temporal artifacts in the resulting low frame rate videos, accompanied by loss in perceptual video quality. The amount of distortion in the low frame rate video depends on the video content and the amount of motion in the video. In this section we investigate the relationship between quality degradation due to frame rate conversion and the level of aliasing in the frequency domain.

The values of a sample pixel in a video over time constitutes a one dimensional signal (Figure 5.1). We refer to this signal as a “pixel signal”. The frequency spectrum of a pixel signal can be generated by applying the Discrete Fourier Transform (DFT). When down-sampling is performed on a pixel signal, the frequency spectrum of the original pixel signal is repeated with a smaller period than the original signal spectrum (Figure 5.2). It can cause overlaps between successive repetitions of the original signal's spectrum. This overlap is termed aliasing. We will use the power of the signal in this aliasing region to estimate the information loss and predict perceptual video quality.

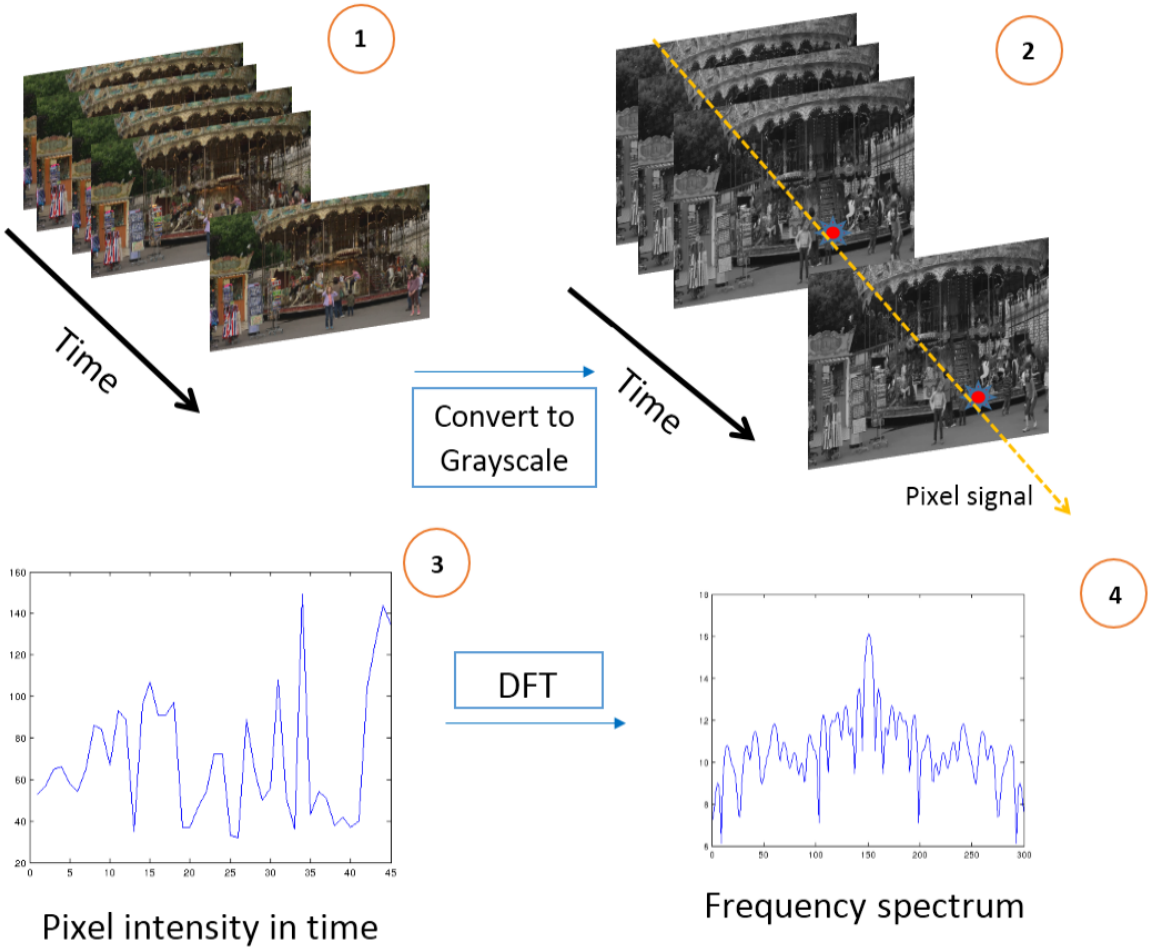


Figure 5.1: An example of pixel intensity in time and the corresponding frequency spectrum.

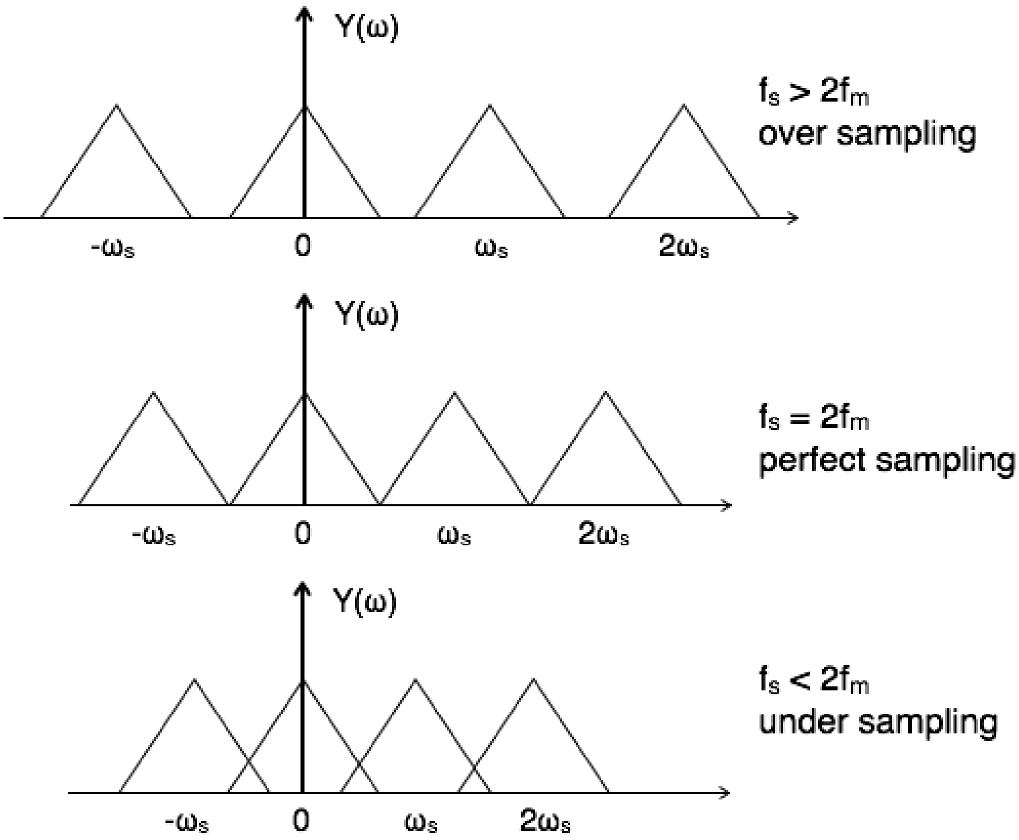


Figure 5.2: Aliasing produced by sampling with frequency lower than the Nyquist rate and without pre-filtering.

In order to find the power in the aliasing region that we refer to as aliasing power, we perform frequency analysis by modeling the pixel signal. We consider a video as a three-dimensional signal $V(r, c, t)$ where r refers to the row, c refers to the column, and t refers to the time component.

Given a pixel in a video frame, we name its row r_i, and its column c_j. Then the one-dimensional pixel signal as a function of time (t) is defined as

$$u_{r_i,c_j}(t) = \{V_g(r,c,t)|r = r_i, c = c_j\}. \tag{5.1}$$

By applying a Fourier transform to the pixel signal the frequency spectrum is obtained by

$$s_{r_i,c_j}(f) = \int u_{r_i,c_j}(t)e^{-j2\pi ft}dt, \tag{5.2}$$

where $s_{r_i,c_j}(f)$ is the frequency spectrum. The frequency spectrum for a sample pixel in one of the videos of the IVC-HFRVQA database and the steps to create the frequency spectrum are illustrated in Figure 5.1.

The pixel signals obtained from a sequence of frames in a video can be viewed as the discrete version of a continuous signal obtained by identifying the pixel value at each instance of time. Due to the limitations of our digital machines, we can only store and process the discrete version of a pixel signal. Therefore, the pixel signal is a digital representation of the continuous pixel signal obtained from a process of sampling. In this sampling process, the frequency spectrum of the discrete signal is a repeated version of the continuous signal's spectrum with a period equaling the sampling frequency, as shown in Figure 5.2. This repetition may cause aliasing between two repetitions of the spectrum. In theory, the pixel signal may have an infinite bandwidth. In practice, it is typically reasonable to assume a signal to be bandlimited. An example of a frequency spectrum for a continuous pixel signal, together with an example of a sampling process with or without aliasing is shown in Figure 5.2. If the sampling frequency is less than the Nyquist rate, aliasing will be present, as indicated by the overlapping areas of the spectrum repetitions. The size of the overlapping area, indicated as the S1 region in Figure 5.3, depends on the sampling frequency and the signal bandwidth. We can use the power of the aliasing part to estimate

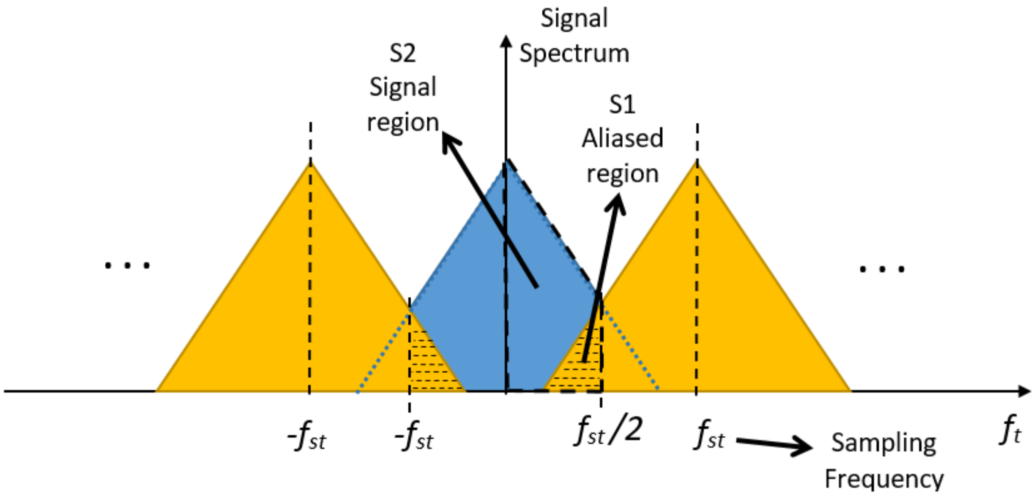


Figure 5.3: Frequency spectrum of sampled signal when aliasing occurs.

the aliasing distortions caused by frame rate reduction or temporal sampling. The power of the signal in the aliased region (S1 in Figure 5.3) is given by

$$P_{r_i,c_j}(f_{st}) = \int_{0}^{f_{st}/2} |s_{r_i,c_j}(f_{st} - f_t)|^2 df_t = \int_{f_{st}/2}^{f_{st}} |s_{r_i,c_j}(f_t)|^2 df_t \tag{5.3}$$

where $|s_{r_i,c_j}(f_t)|$ is the magnitude of the frequency component in the frequency spectrum of the pixel signal for row r_i and column c_j at temporal frequency of f_t, f_{st} is the sampling frequency and $P(f_{st})$ is the estimate of aliasing power.

A more meaningful way to judge the impact of aliasing is by measuring aliasing power relative to the power distribution of the entire signal. As it can be seen in Figure 5.3, when aliasing occurs, the aliasing region is influenced by another period of frequency spectrum repetition. To measure the power in the aliased region relative to the signal power (S2 region in Figure 5.3), we define a normalized aliasing factor by

$$A_{T,r_i,c_j}(f_{st}) = \frac{\int\limits_{0}^{f_{st}/2} |s_{r_i,c_j}(f_{st}-f_t)|^2 df_t}{\int\limits_{0}^{f_{st}/2} |s_{r_i,c_j}(f_t)|^2 df_t} = \frac{\int\limits_{f_{st}/2}^{f_{st}} |s_{r_i,c_j}(f_t)|^2 df_t}{\int\limits_{0}^{f_{st}/2} |s_{r_i,c_j}(f_t)|^2 df_t}. \tag{5.4}$$

where the power of the signal in the aliased region is normalized by the power of the signal in the low frequency region of the spectrum. This aliasing factor can be used as an estimate of aliasing strength by frame rate down-sampling.

The temporal aliasing factor is computed for each pixel signal (as in Figure 5.1) extracted from the video and averaged to yield an overall temporal aliasing factor of the sampled video by

$$A_T(f_{st}) = \frac{1}{N \times M} \sum_{i=1}^{N} \sum_{j=1}^{M} A_{T,r_i,c_j}(f_{st}). \tag{5.5}$$

where N is the number of pixels in a frame and $A_{T,r_i,c_j}(f_{st}$ is the temporal aliasing factor for pixel locate at row r_i and column c_j. Considering the local similarity of pixel values in the video frame, for computational efficiency, a random subset of pixel signals may be used to represent all pixel signals. Using the subset of pixels in the calculation of Eq. 5.25, the estimated temporal aliasing factor is computed by

$$\widetilde{A}_T(f_{st}) = \frac{1}{k} \sum_{(r_i,c_j)\in\phi} A_{T,r_i,c_j}(f_{st}). \tag{5.6}$$

where ϕ is the set of the selected pixels in video and k is the number of pixels in this set.

For videos in the IVC-HFRVQA-I database with different frame rates which are constructed from the original 60fps, we assume that 60fps is a sampling rate without aliasing and lower frame rates may result in aliasing due to the overlaps in the frequency spectrum.

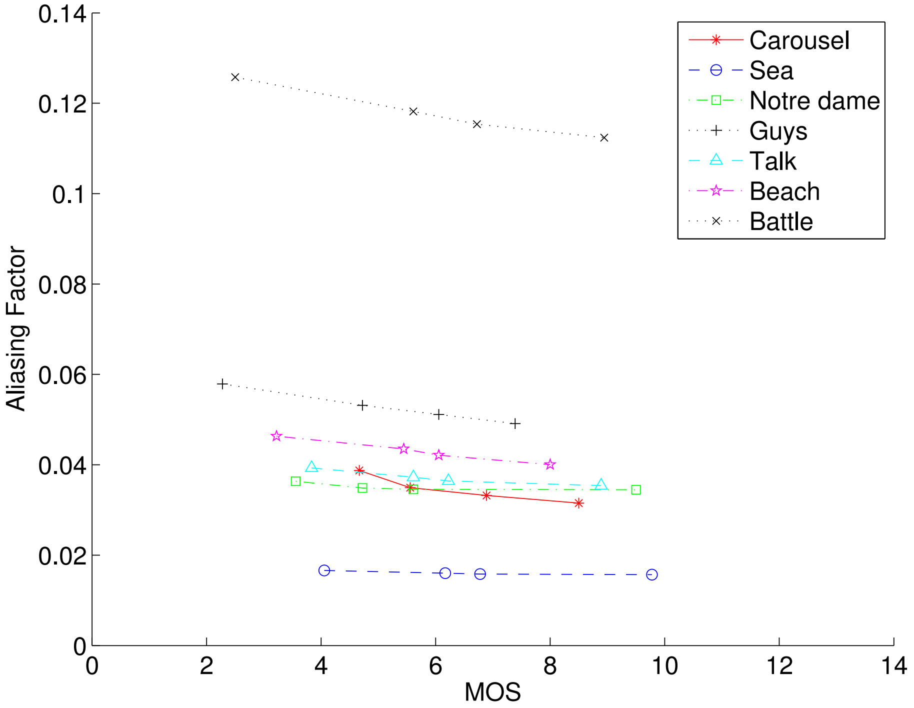


Figure 5.4: Estimated temporal aliasing factor $(\widetilde{A}_T)$ vs MOS for four different down-sampling rate in IVC-HFRVQA-I database.

Figure 5.4 shows the temporal aliasing factor $(\widetilde{A}_T)$ for different video content in the IVC-HFRVQA-I database for four different frame rates of 30, 15, 10, and 5fps, respectively, created from the original 60fps videos.

5.1.2 2D Spatiotemporal Aliasing Factor (X-T and Y-T)

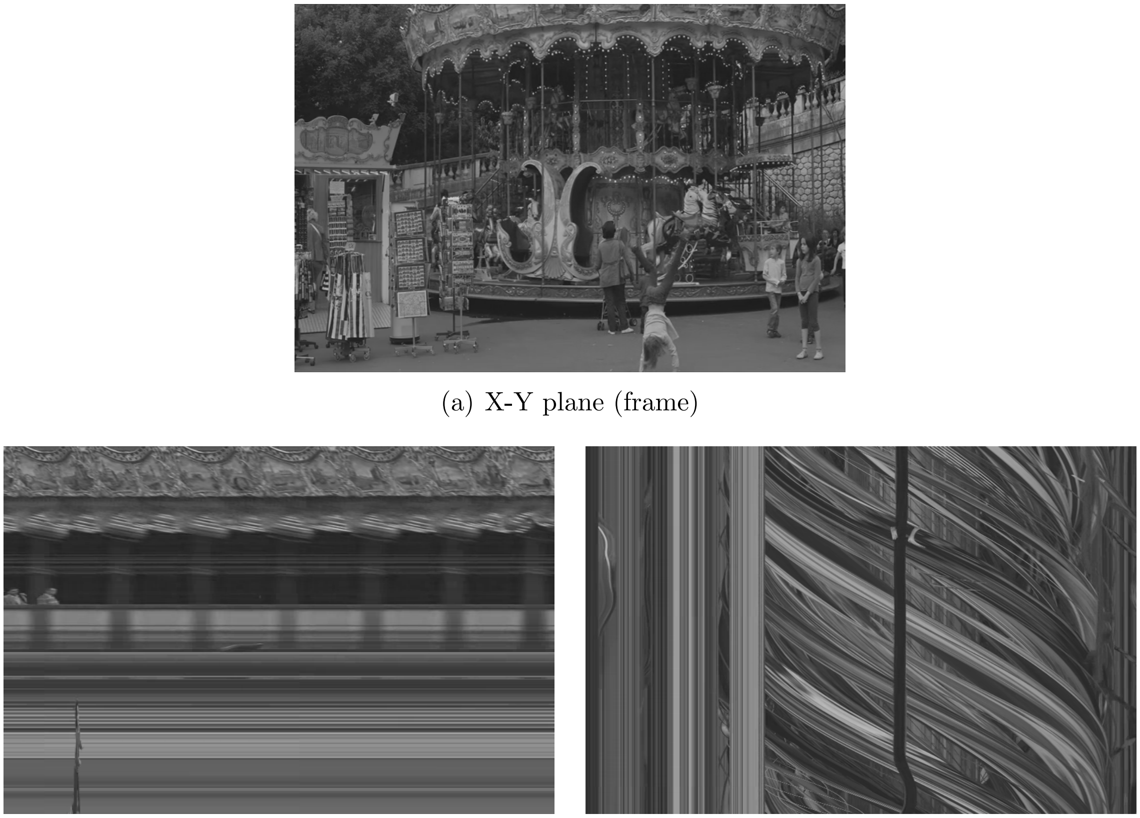
In the previous section, we analyzed the effect of aliasing along the temporal direction. The temporal analysis of pixel signals does not consider the spatial context. For example the shifts of pixel values in the horizontal or vertical directions in consecutive frames are not taken into account. In this section, we extend our analysis to joint spatio-temporal analysis.

A simple model of spatio-temporal analysis is time-space plane analysis. The joint observation of space and time is performed on the two-dimensional plane of X-T or Y-T for horizontal or vertical analysis respectively in which 'X' represents a row, 'Y' represents a column, and 'T' indicates the time axis. Geometrically, we can observe these planes as an intersection of the horizontal/vertical plane with the video sequence in a three-dimensional spatio-temporal space. Figure 5.5 shows two sample of spatio-temporal planes extracted from a video in the IVC-HFRVQA-I database. It can be observed that the X-T and Y-T planes create very different pattern from the spatial X-Y plane and provide a different perspective on horizontal and vertical motion. Non-smooth motion due to frame dropping can be captured in these planes better than pixel signals.

Using the X-T (or Y-T) plane, and the corresponding two dimensional frequency spectrum $s(f_t, f_x)$, the aliasing due to temporal down-sampling is indicated as the overlap of frequency spectra with its repetition in the temporal direction (f_t) as shown in Figure 5.6. This overlapping can also occur due to spatial down-sampling, which corresponds to frame resolution reduction in video. Considering the general case when both temporal and spatial down-sampling may be applied to a video, the aliasing region may happen in both directions in frequency spectra as shown in Figure 5.7. Therefore, a spatiotemporal aliasing factor is calculated by

$$A_{XT,r_i}(f_{sx}, f_{st}) = \frac{\int\limits_{0}^{f_{sx}/2} \int\limits_{0}^{f_{st}/2} |s_{r_i}(f_{st} - f_t, f_{sx} - f_x)|^2 df_t df_x}{\int\limits_{0}^{f_{sx}/2} \int\limits_{0}^{f_{st}/2} |s_{r_i}(f_t, f_x)|^2 df_t df_x}. \tag{5.7}$$

where r_i is the index of a selected row to create the X-T plane, f_{st} is the sampling frequency in temporal direction, and f_{sx} is the sampling frequency in X direction. A similar method may be applied on a Y-T analysis for a selected column to calculate $A_{YT,c_j}(f_{sy}, f_{st})$ for f_{sy}



(a) X-Y plane (frame)

(b) Y-T plane

(c) X-T plane

Figure 5.5: X-Y, X-T, and Y-T planes constructed from 3D video volume [3].

to be sampling frequency in Y direction. This aliasing factor is calculated for each X-T plane (corresponding to each row r_i) and Y-T plane (corresponding to each column c_j) in the sequence of video frames. The average of this spatio-temporal aliasing factor for all X-T planes results in the overall 2D spatiotemporal aliasing factor (A_{XT}) for the entire video and is calculated as follows

$$A_{XT}(f_{st}) = \frac{1}{N}\sum_{i=1}^{N} A_{XT,r_i}(f_{st}), \tag{5.8}$$

where N is the number of rows in the video frame and $A_{XT,r_i}(f_{st})$ is the 2D spatiotemporal

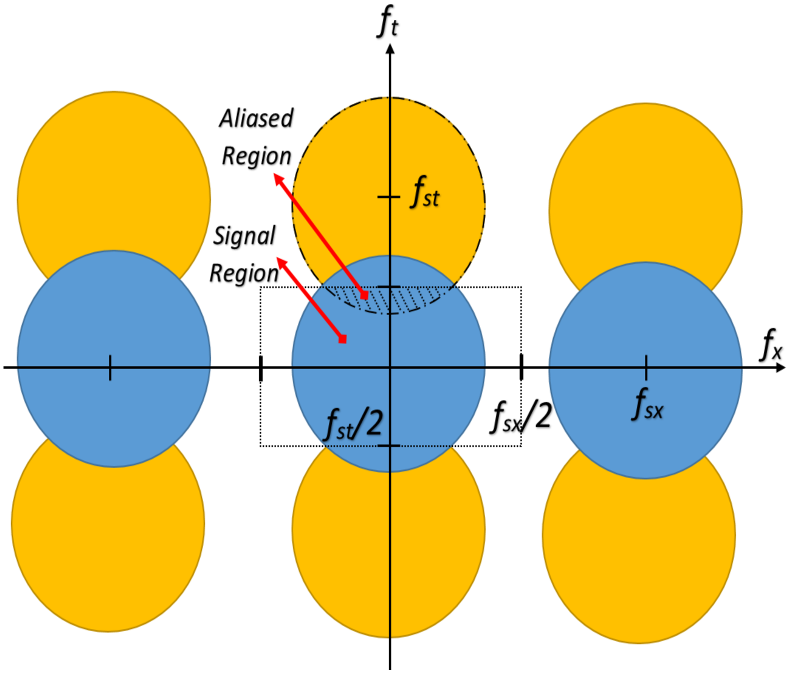


Figure 5.6: Aliasing in temporal down-sampling in two-dimensional frequency spectrum of X-T plane is represented by the overlapping region between central spectrum and its repetitions when the sampling rate is lower than the Nyquist rate.

aliasing for an X-T plane corresponding to r_i. The same approach can be used to define 2D spatiotemporal aliasing factor for Y-T planes created from columns (A_YT).

$$A_{YT}(f_{st}) = \frac{1}{M}\sum_{j=1}^{M} A_{YT,c_j}(f_{st}), \tag{5.9}$$

where M is the number of columns in the video frame. Similar to the temporal aliasing factor calculation, the computational costs may be reduced by performing calculations on selected lines (rows or columns) using

$$\widetilde{A}_{XT}(f_{st}) = \frac{1}{R}\sum_{r_i \in \Psi} A_{XT,r_i}(f_{st}) \tag{5.10}$$

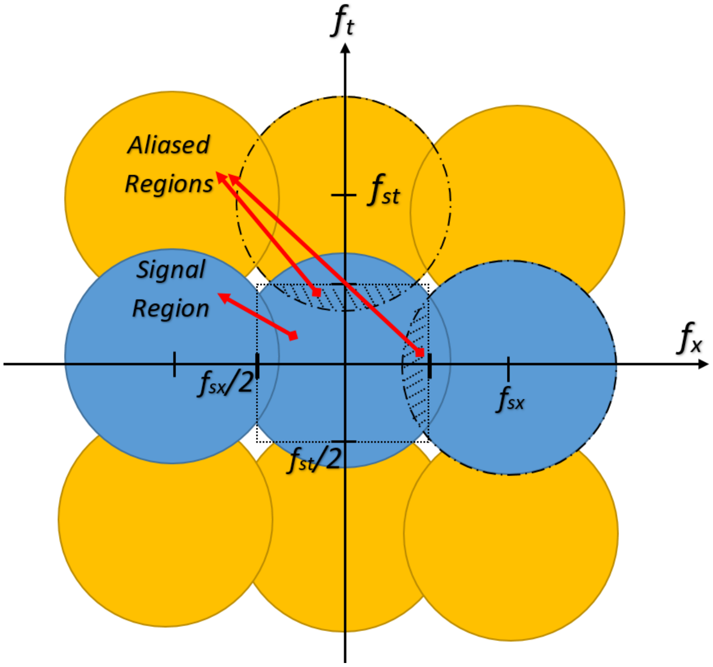


Figure 5.7: Aliasing in spatial and temporal down-sampling in two-dimensional frequency spectrum of X-T plane is represented by the overlapping region between central spectrum and its repetitions when the sampling rate is lower than the Nyquist rate.

$$\widetilde{A}_{YT}(f_{st}) = \frac{1}{C} \sum_{c_j \in \Omega} A_{YT,c_j}(f_{st}), \tag{5.11}$$

where Ψ is a selected subset of rows and Ω is a selected subset of columns in a frame, and R and C are the sizes of these subsets, respectively.

5.1.3 2D Spatial Aliasing Factor

Similar to the analysis of X-T and Y-T planes explained in the previous chapter, 2D aliasing can be performed on the 2D X-Y plane in which the aliasing will be used as an estimation of information loss due to frame size reduction. Considering the X-Y plane in Figure 5.5

and its corresponding frequency spectrum, the 2D spatial aliasing factor can be calculated using

$$A_{XY,f_i}(f_{sx}, f_{sy}) = \frac{\int\limits_0^{f_{sx}/2} \int\limits_0^{f_{sy}/2} |s(f_{sx} - f_x, f_{sy} - f_y)|^2 df_x df_y}{\int\limits_0^{f_{sx}/2} \int\limits_0^{f_{sy}/2} |s(f_x, f_y)|^2 df_x df_y}. \tag{5.12}$$

where f_{sx} and f_{sy} are the sampling frequency in spatial direction of X and Y respectively.

5.1.4 3D Spatio-temporal Aliasing

The aliasing analysis of digital videos can be performed in the complete three-dimensional space of XYT. We consider the power of the signal in the 3D overlapped volume in 3D frequency space analysis. The 3D spatio-temporal aliasing factor is then given by

$$A_{XYT}(f_{sx}, f_{sy}, f_{st}) = \frac{\int\limits_0^{f_{sx}/2} \int\limits_0^{f_{sy}/2} \int\limits_0^{f_{st}/2} |s(f_{st} - f_t, f_{sx} - f_x, f_{sy} - f_y)|^2 df_t df_x df_y}{\int\limits_0^{f_{sx}/2} \int\limits_0^{f_{sy}/2} \int\limits_0^{f_{st}/2} |s(f_t, f_x, f_y)|^2 df_t df_x df_y}. \tag{5.13}$$

5.1.5 Perceptual Contrast Sensitivity

The aliasing factors introduced in the previous sections assume the same importance for all frequency components, but human visual perception has different sensitivity to different frequencies [4, 84]. This sensitivity is characterized by the visual Constrast Sensitivity Function (CSF) which is defined as a function of visual sensitivity in terms of both temporal

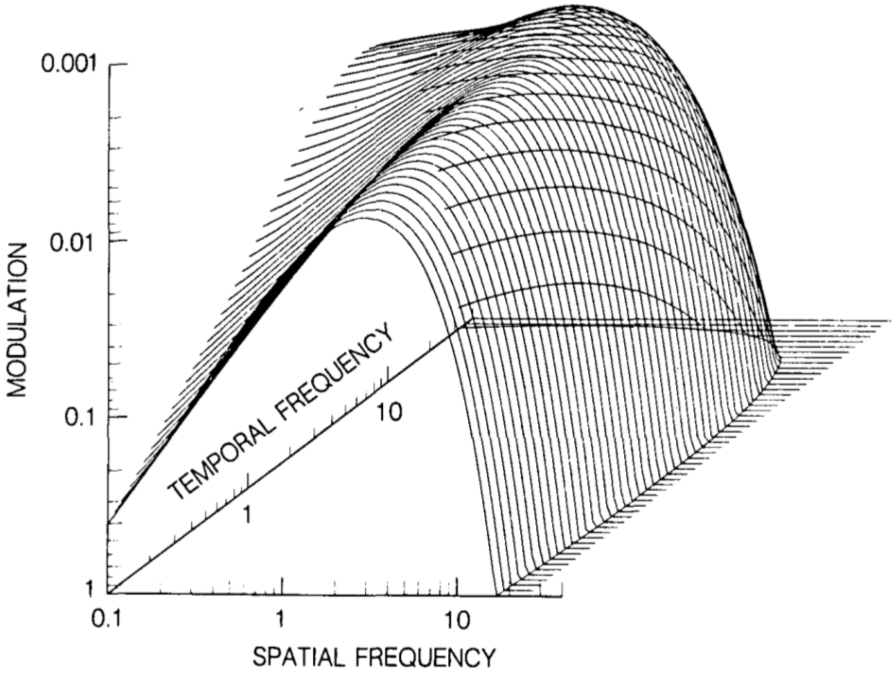


Figure 5.8: Spatiotemporal contrast sensitivity function as function of both temporal and spatial frequency components [4].

and spatial frequencies. Kelly explored the CSF for moving pictures by psycho-visual experiments with stimuli of different spatial and temporal frequencies [4] and modeled the CSF function as a surface on spatial and temporal frequency space (Fig. 5.8). This function has been used in many subsequent studies [84, 85] and quantified by [84] as

$$\lambda(f, v_r) = kc_0c_2v_R(c_12\pi f)^2 exp(\frac{-c_14\pi f}{f_{max}}) \tag{5.14}$$

where k and f_{max} are defined as

$$k = s_1 + s_2|log(\frac{c_2v_R}{3})|^3, f_{max} = \frac{f_1}{c_2v_R + 2}, \tag{5.15}$$

where, $s_1 = 6$, $s_2 = 7.3$, $f_1 = 45.9$, $c_0 = 1.14$, $c1 = 0.67$, and $c_2 = 1.92$ are constants selected according to [84]. v_r is the retinal velocity and f is the spatial frequency. λ is the sensitivity as a function of f and v_r. Following the X-T analysis and considering motions projected in X-T planes, spatial frequency in Eq. (5.14) can be estimated by the spatial frequency of X-T plane (f_x) by using

$$f \approx g(f_x) = f_x\ D, \tag{5.16}$$

where D is the angular resolution measured by pixel/degree unit. The same analysis can be defined for Y-T planes. Retinal velocity (v_r) can be estimated by spatial and temporal frequency components by

$$v_R \approx h(f_t, f_x) = \frac{f_t\ R}{f_x} \tag{5.17}$$

where, R is the frame rate. Using Eq. (5.16), (5.17), we obtain an estimate of the sensitivity function (λ) as a function of f_t and f_x as follows

$$\widetilde{\lambda}(f_t, f_x) = \lambda(g(f_x), h(f_t, f_x)). \tag{5.18}$$

where $\widetilde{\lambda}$ is the estimate of the sensitivity function for X-T plane frequency analysis. The same equation can be applied to f_y for Y-T planes and define $\widetilde{\lambda}(f_t, f_y)$.

The sensitivity function for 1D signals of T, X, and Y can be extracted from two-dimensional λ by fixing the second input parameter to a typical average frequency value of the video based on the size and content. These sensitivity functions are applied in the proposed aliasing factors to consider human visual perception characteristics. The aliasing factors proposed in previous sections can be modified to perceptual factors by applying CSF-based weighting.

5.1.6 Perceptual Temporal Aliasing Factor

We extract a simplified 1D temporal contrast sensitivity function from 2D sensitivity function (λ) in the previous section and apply it to the pixel signal in temporal aliasing analysis. To calculate 1D temporal contrast sensitivity, the spatial frequency component (f_x) in Eq.5.18 is fixed to a specific value as follows

$$\widetilde{\lambda}_t(f_t) = \widetilde{\lambda}(f_t, f_{x_0}) = SF(g(f_{x_0}), h(f_t, f_{x_0})). \tag{5.19}$$

where the value of f_{x_0} is the fixed spatial frequency, selected to be one quarter of the maximum possible frequency in a video frame by considering all frame resolutions, display size, and viewing distance of observer. We selected one quarter of maximum possible spatial frequency as the spatial frequency in frames are usually much lower than the maximum which correspond to the pattern of the black and while neighboring pixels. We considered at least 4 pixel for the details of frame which. The selected fixed spatial frequency is close to the resolution considered in pixel-wise operators normally defined and used in basic image processing operations.

Using the definition in (5.19), the perceptual temporal aliasing factor is defined based on (5.25) by

$$A_{PT,r_i,c_j}(f_{st}) = \frac{\int\limits_0^{f_{st}/2} \lambda_t(f_t)|s_{r_i,c_j}(f_{st} - f_t)|^2 df_t}{\int\limits_0^{f_{st}/2} \lambda_t(f_t)|s_{r_i,c_j}(f_t)|^2 df_t}. \tag{5.20}$$

The perceptual temporal aliasing factor for the entire video is defined in a similar way by pooling from the selected pixels as

$$\widetilde{A}_{PT}(f_{st}) = \frac{1}{k} \sum_{(r_i,c_j)\in\phi} A_{PT,r_i,c_j}(f_{st}). \tag{5.21}$$

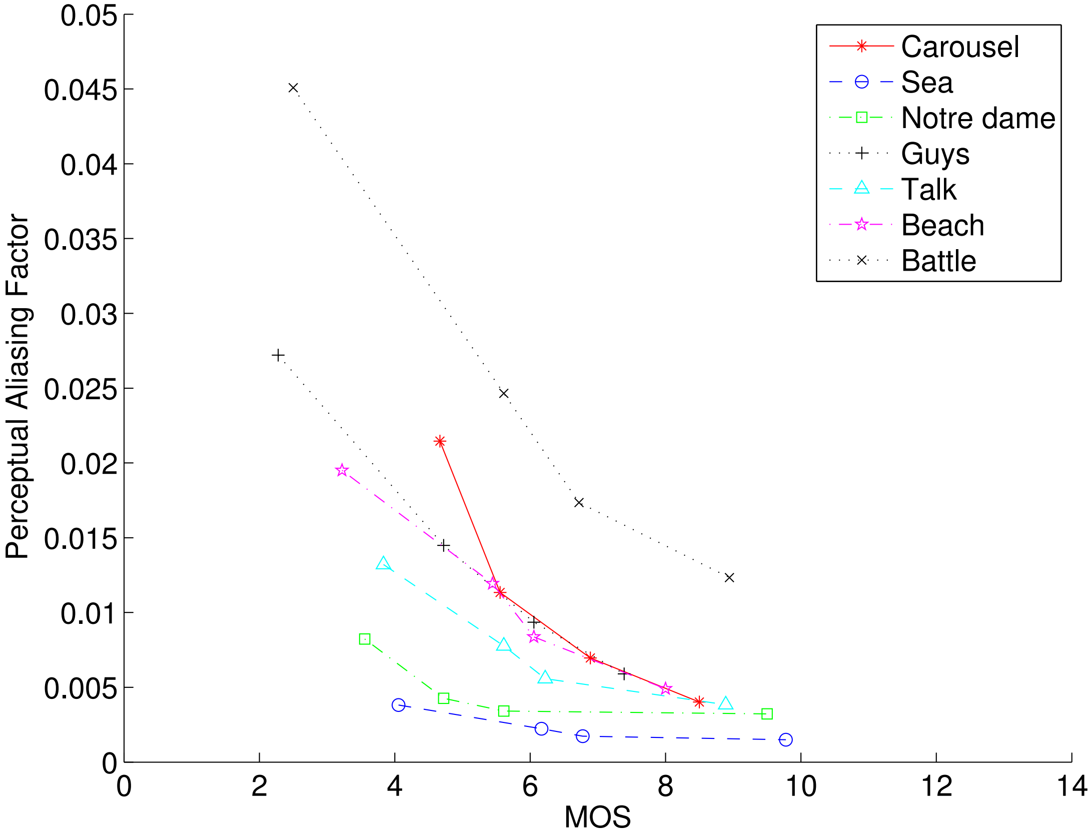


Figure 5.9: Perceptual aliasing factor versus MOS for video sequences at different frame rates without compression.

The perceptual temporal aliasing factors $A_{PT}(f_{st})$ of the videos in IVC-HFRVQA-I database are calculated using. (5.21) and the results are reported in Figure 5.9 for all seven contents and four levels of frame rates, by assuming no aliasing in the 60fps videos in the database. It can be observed that the perceptual aliasing factor decreases with the frame rate for each individual sequence, but the rate of decrement varies. It is also interesting to see that for high motion or high texture videos, the decreasing trend is more significant. This may be because the high motion videos or those videos with complex content but moderate motion have significant energy in the high frequency components of the spectrum.

5.1.7 Perceptual 2D Spatio-Temporal Aliasing Factor

The sensitivity function (λ) may be incorporated with the proposed two dimensional spatiotemporal aliasing factors. Using two dimensional sensitivity as defined in (5.18) as the weighting factor, the perceptual X-T aliasing factor by using (5.7) is defined as follows

$$A_{PXT,r_i}(f_{xt}) = \frac{\int_0^{f_{sx}/2} \int_0^{f_{st}/2} \widetilde{\lambda}(f_t, f_x) |s(f_{st} - f_t, f_x)|^2 df_t df_x}{\int_0^{f_{sx}/2} \int_0^{f_{st}/2} \widetilde{\lambda}(f_t, f_x) |s(f_t, f_x)|^2 df_t df_x} \tag{5.22}$$

$$\widetilde{A}_{PXT}(f_{st}) = \frac{1}{R} \sum_{r_i \in \Psi} A_{PXT,r_i}(f_{st}) \tag{5.23}$$

A similar method applies to Y-T planes. The calculated perceptual aliasing factor on uncompressed videos of the IVC-HFRVQA-I database is shown in the Figure 5.10, where it can be seen that the correlation between the aliasing factor and MOS for different frame rates of various contents are better than the pixel signal frequency analysis in the previous section.

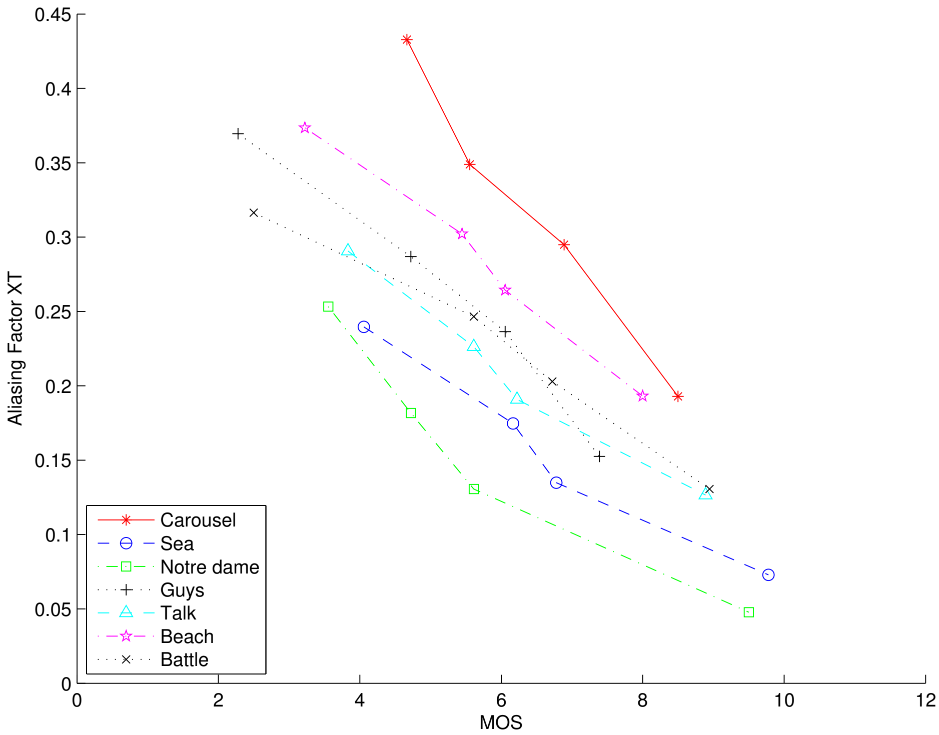


Figure 5.10: X-T aliasing factor versus MOS for videos in IVC-HFRVQA-I database.

5.1.8 Perceptual Spatial Aliasing Analysis

Resolution change is a down-sampling of video signals in the spatial dimensions. The 1D temporal aliasing can be rewritten for 1D-X signal (for rows) and 1D-Y signals (for columns) on selected rows and columns of video frames. For perceptual aliasing analysis, $\widetilde{\lambda}(f_x)$ and $\widetilde{\lambda}(f_y)$ can be extracted from the original definition of $\widetilde{\lambda}$ by fixing f_t to a selected value:

$$A_{PX,r_i,c_j}(f_{st}) = \frac{\int\limits_{0}^{f_{sx}/2} |s_{f_i,y_j}(f_{sx}-f_x)|^2 df_x}{\int\limits_{0}^{f_{sx}/2} |s_{f_i,y_j}(f_x)|^2 df_x}. \tag{5.24}$$

$$A_{PX}(f_{sx}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} A_{PX,f_i,c_j}(f_{sx}). \tag{5.25}$$

Considering the 2D case of X-Y plane, we calculate the perceptual 2D spatial aliasing as

$$A_{PXY,f_i}(f_{sx}, f_{sy}) = \frac{\int\limits_0^{f_{sx}/2}\int\limits_0^{f_{sy}/2} \widetilde{\lambda}(f_x)\widetilde{\lambda}(f_y)|s(f_{sx}-f_x, f_{sy}-f_y)|^2 df_x df_y}{\int\limits_0^{f_{sx}/2}\int\limits_0^{f_{sy}/2} \widetilde{\lambda}(f_x)\widetilde{\lambda}(f_y)|s(f_x, f_y)|^2 df_x df_y} \tag{5.26}$$

$$\widetilde{A}_{PXY}(f_{sx}, f_{sy}) = \frac{1}{F}\sum_{f_i\in\phi} A_{PXY,f_i}(f_{sx}, f_{sy}) \tag{5.27}$$

where f_{sx} and f_{sy} are the sampling frequency in the x and y dimensions, respectively, and F is the number of frames selected for the analysis from ϕ set. For perceptual analysis, the original sensitivity function is converted to one-dimensional spatial sensitivity functions ($\lambda(f_x)$ and $\lambda(f_y)$) in the same way by fixing f_t to one quarter of the maximum possible temporal frequency.

5.1.9 Perceptual 3D Spatio-Temporal Aliasing Factor

The perceptual analysis can be extended to the three dimensional frequency analysis in XYT space. The two dimensional sensitivity function of λ is extended by introducing spatial frequency component (f_{xy}) as follows

$$f_{xy} = \sqrt{f_x^2 + f_y^2}. \tag{5.28}$$

f_{xy} integrates X and Y frequency components by considering the distance of frequency component to the origin at X-Y frequency spectrum. Using f_{xy} in the original 2D equation in (5.18), the sensitivity function for XYT analysis is defined as

$$\widetilde{\lambda}(f_t, f_x, f_y) = \lambda(g(f_{xy}), h(f_t, f_{xy})). \tag{5.29}$$

The equations for perceptual 2D aliasing factors introduced in the previous sections can be rewritten for the 3D cases by using $\lambda(f_t, f_x, f_y)$ to calculate $\widetilde{A}_{PXYT}(f_{st}, f_{sx}, f_{sy})$.

5.2 Objective Video Quality Assessment Incorporating Aliasing Factors

5.2.1 Video Quality Assessment Framework

To design a comprehensive objective VQA method, the entire path from acquisition of the pristine source video to the display at the end user side should be considered. Figure 5.11 shows a typical path of video encoding, transmission, and display which contains a series of processing steps. Each step in this flow may introduce artifacts and cause quality degradations. Ignoring the device dependent post processing at the end users' side, the video quality control is usually performed by controlling the encoding parameters on the encoding side before channel transmission as shown in Figure 5.11, where the quality might be affected by the frame rate change, the resolution change, the compression level, or their combinations. Therefore the entire quality prediction model (Q) can be described as a function of the frame rate down-sampling factor, the resolution down-sampling factor, and

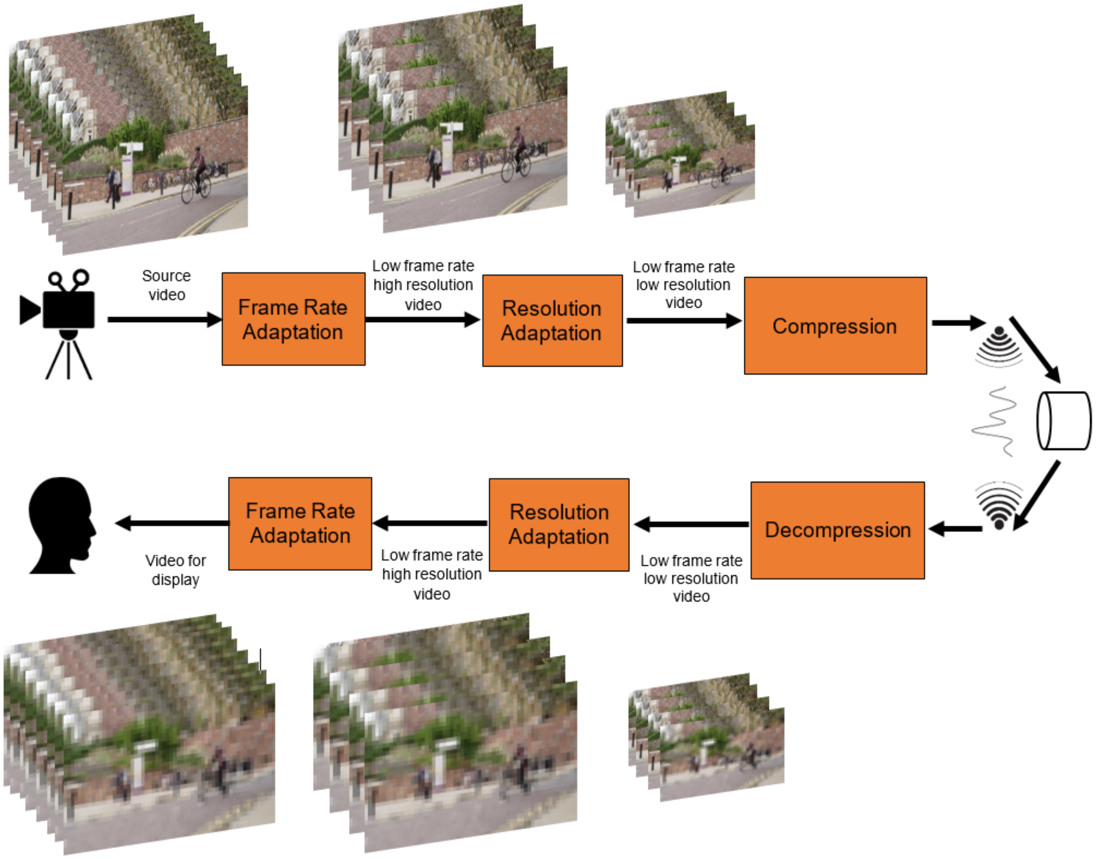

Source video
Frame Rate Adaptation
Low frame rate high resolution video
Resolution Adaptation
Low frame rate low resolution video
Compression
Decompression
Low frame rate low resolution video
Resolution Adaptation
Low frame rate high resolution video
Frame Rate Adaptation
Video for display

Figure 5.11: Video processing path from source video acquisition to display at end user side.

the compression ratio as follows

$$Q_E = Q(\Delta_{fr}, \Delta_{res}, r_{comp}) \tag{5.30}$$

where Δ_{fr}, and Δ_{res} are down-sampling factors along temporal and spatial directions, and r_{comp} is the compression ratio. The frame rate and resolution change blind VQA methods discussed in Chapter 2, are only supporting quality modeling by considering r_{comp}.

The aliasing factors for proposed in the previous section were targeted at capturing the quality degradations due to frame rate and resolution changes change (considering Δ_{fr} and Δ_{res}).

$$Q_E^{Alias}(\Delta_{fr}, \Delta_{res}) \tag{5.31}$$

However, in real applications any combination of resolution change, frame rate change, and compression may occur. Therefore, to design a comprehensive objective VQA model, the aliasing factor are not sufficient, and the impact of compression algorithms on quality should also be taken into account.

Quality degradation due to lossy compression usually happens because of the quantization of frequency components. The quantization effect, when combined with the impact of motion estimation/compensation and block-based encoding, causes visible artifacts such as blocking and blurring. Based on the compression algorithms used in encoding and the level of compression selected in the encoding process, different types and numbers of artifacts may appear in the compressed video.

Quality degradation due to compression has been the topic of many earlier works and some VQA methods are widely recognized in the literature. The quality predictive model considering compression artifacts only as a simplified version of Eq. 5.30 can be expressedk as

$$Q_E^{Comp}(r_{comp}) \tag{5.32}$$

where r_{comp} is the compression ratio factor. These VQA methods are blind to frame rate and resolution changes (Δ_{fr} and Δ_{res}). We opt to use a frame rate and resolution change blind VQA method in combination with our proposed aliasing factors to design a comprehensive VQA method that could perform quality assessment task in any combination of changes. From Eq. 5.30, we can express this comprehensive model by using Eq. 5.31 and Eq. 5.32

as

$$\tilde{Q_E} = \Psi(Q_E^{Alias}, Q_E^{Comp}), \tag{5.33}$$

where Ψ is a function that integrates these two parameters. For compression level quality assessment method, we specifically selected MS-SSIM, which is one of the most well-known VQA methods that has shown high quality prediction performance and computational efficiency in predicting the quality of compressed videos [23]. It is also the best performing baseline model according to our test in Section 3.3.

Figure 5.12 shows the schematic diagram of the proposed method which uses MS-SSIM and a selection of aliasing factors as input. For the integration method, Ψ function in Eq. 5.33, we used Support Vector Regressor to predict video quality given two types of inputs (aliasing factor(s) and MS-SSIM scores). We select Support Vector Regression (SVR) because it is a simple and robust learning algorithm that works well in complex training tasks.

As one of the input parameter in the proposed method, the aliasing factors were calculated by analyzing the frequency spectrum of the high resolution and high frame rate source video of each content in the database. For each compressed video (test video) in the dataset , the corresponding version of the video content with the frame rate of 120 fps and the resolution of 1920×1080 and with lossless compression (qp=0) is selected as the reference video. Given a the reference video of a test video, the aliasing factors are calculated by using frequency spectrum of the reference video together with the frame rate and resolution down-sampling ratios to determine the sampling frequency in Eqs. 5.20, 5.22, and 5.26.

For calculating MS-SSIM scores, we compared the video before and after applying compression in Figure 5.11 to take into account the impact of compression. We consider the source video after applying resolution change and frame rate change as a reference for

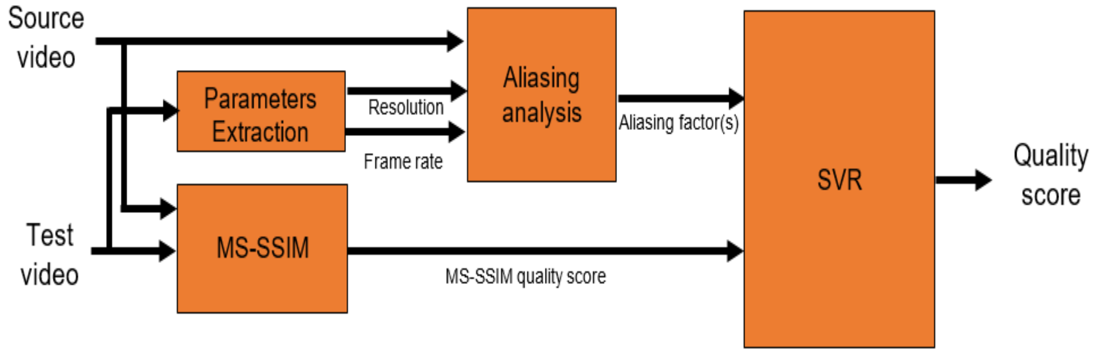


Figure 5.12: Schematic diagram of the proposed VQA method based on aliasing factor.

MS-SSIM computation when evaluating a compressed video. We used the implementation provided by the authors of the original paper [23] as provided in [86].

The SVR is trained by videos in IVC-HFRVQA-II database produced using different combinations of encoding parameters. For each video in the dataset, we used frame rate and resolution change of the test video to calculate the aliasing factors and we calculate MS-SSIM as described before as inputs to SVR. Extracting the SVR's input parameters for all videos in the dataset, We used leave-one-out cross validation method for training and testing. Specifically, we select the videos in the database from all content except one for training the SVR, and use the remaining content for testing. This process is repeated for all possible combinations of training and testing combinations, and the final reported performance measures are the average of all repetitions.

5.2.2 VQA Methods for Comparison

We compare the performance of the proposed methods with state-of-the-art VQA methods. We use eight objective VQA models for performance evaluation including MSE, PSNR, SSIM [22], MS-SSIM [23], VQM [87], VMAF [62], Q-STAR [88], and FRQM [61].

Since none of these methods except for Q-STAR supports cross-resolution video quality assessment, the test videos were up-sampled to 1920×1080 resolution for evaluations. For those methods that do not support cross-frame rate video quality evaluation, including MSE, PSNR, SSIM, MS-SSIM, and VQM, the test videos were up-sampled to 120 fps in temporal direction by repeating frames in the lower frame rate videos. From the cross-frame rate VQA methods reviewed in Chapter 2, we used Q-STAR, FRQM, AND VFAM for comparison. Other methods are not selected for comparisons, because some of them use similar approaches to the ones selected, and some are designed for the low bit-rate low-resolution analysis, or for specific cases of 3D video analysis only.

SSIM has been implemented in many libraries and mathematical tools such as MATLAB, OpenCV and R and we used its implementation in MATLAB. For MS-SSIM, the implementation provided by the authors of the original paper [23] was used as provided in [86]. The code provided by ITS (Institute for Telecommunication Science) as in [87] was used with default parameter settings. For VMAF, the implementation is publicly available as a script [89]. The Python implementation of VMAF is used with default parameters. The FRQM [61] was tested using the code by the main author of the original paper with the provided settings of parameters. As FRQM does not support resolution change, for the quality assessment of any test video, the pristine reference was converted to the same resolution of frames first (with bi-cubic interpolation) and then passed to the code for quality evaluation. For the Q-STAR methods, the parameters proposed in the main paper [88] were used. We used the quality score (MOS) value of the reference video as reference in the implementation of Q-STAR as suggested by the original paper.

5.2.3 Performance in Quality Prediction

The list of all spatial and temporal aliasing factors proposed in this research is reported in Table 5.1. To exploit the potentials of the spatial, temporal, and spatiotemporal aliasing factors in quality prediction, different combination of aliasing factors in Table 5.1 are used. We created 7 combinations of the aliasing factors in addition to MS-SSIM to each combination as the inputs for the training algorithm. The selected aliasing factor combinations are reported in Table 5.2. The combinations are selected to cover different variations of 1D, 2D, 3D aliasing factors and also cover variations of spatial and temporal aliasing to investigate the effect of the proposed aliasing factors in predicting video quality. Each selected set of aliasing factors is combined with the MS-SSIM score to form a complete feature vector as the input to train the SVR, as explained in the previous section.

Table 5.1: List of aliasing factors used for training.

Aliasing factor	Description
A_{PT}	1D temporal aliasing factor defined on pixel signals
A_{PX}	1D spatial aliasing factor defined on rows of frames
A_{PY}	1D spatial aliasing factor defined on columns of frames
A_{PXT}	2D spatio-temporal alising factor defined on X-T palnes
A_{PYT}	2D spatio-temporal aliasing factor defined on Y-T palnes
A_{PXY}	2D spatial aliasing factor defined on frames
A_{PXYT}	3D spatio-temporal aliasing factor defined on volume video frames

Table 5.2: The selected aliasing factor combination.

Feature set	A_{PT}	A_{PX}	A_{PY}	A_{PXT}	A_{PYT}	A_{PXY}	A_{PXYT}	MS-SSIM
A								x
B	x							x
C						x		x
D							x	x
E	x	x	x					x
F	x					x		x
G				x	x	x		x

The predicted quality values by seven variations of the proposed algorithm are compared against the MOS values from the subjective test in IVC-HFRVQA-II dataset. The

Table 5.3: Performance comparison of VQA methods on IVC-HFRVQA-II dataset.

Method	SRCC	PLCC	RMSE
MSE	0.2437	0.1415	43.65
PSNR	0.2438	0.3500	45.87
SSIM	0.3298	0.2907	33.27
MS-SSIM [23]	0.6423	0.6704	39.03
Q-STAR [88]	0.2654	0.2754	32.20
FRQM [61]	0.4583	0.4012	41.88
VQM [87]	0.2089	0.2501	30.60
VMAF [62]	0.6094	0.5777	32.95
Proposed A	0.7063	0.7212	27.95
Proposed B	0.7352	0.7183	20.77
Proposed C	0.9318	0.9362	9.1889
Proposed D	0.7609	0.7597	21.68
Proposed E	0.9454	0.9418	10.0
Proposed F	0.9592	0.9597	7.8139
Proposed G	0.9650	0.9652	7.20

performance metrics, including PLCC, SRCC, and RMSE, are used to evaluate the proposed algorithm.

Table 5.3 reports the performance evaluation results on the IVC-HFRVQA-II database for seven proposed methods and VQA methods used for comparisons as mentioned in the previous section. Figures 5.13, 5.14, and 5.15 show the scatter plots of predicted values vs. MOS for different variation of proposed VQA methods alongside the selected VQA methods for comparison. Figures 5.16, 5.17 and 5.18 report the comparison of MOS and the predicted values color coded for different frame rates, resolutions and the sequences to provide insights on cross-frame rate , cross-resolution, and cross-content performance.

From the results reported in Table 5.3, we observe that well-known VQA methods such as SSIM, MS-SSIM, and VQM do not perform well in predicting video quality when the frame rate changes alongside other parameters of video such as resolution change and compression level. By contrast, the proposed methods based on aliasing factors shows better performance in general. According the reported performance in Table 5.3 and Figures 5.16, 5.17 and 5.18 from different combinations of aliasing factors used in learning SVR, the combination of spatial and temporal aliasing factors can provide better performance in the quality assessment task. For example the combinations of features in Proposed F (MS-SSIM, XY, T) and Proposed G (MS-SSIM, XT, YT, and XY) show the highest correlations with MOS values where T represents temporal aliasing factor, XT, YT, and XY represent 2D spatial and spatio-temporal aliasing factors calculated in X-T planes (extracted from rows), Y-T planes (extracted from columns), and X-Y planes (for selected frames), respectively.

Comparing the performance of spatial aliasing with temporal aliasing factors shows that the resolution change in general has a greater impact on the quality prediction. This is in direct relationship with the higher impact of resolution in subjective quality assessment reported in Chapter 3. This can be clearly observed from the comparison of Proposed C (MS-SSIM, XY) and Proposed B (MS-SSIM, T) in Table 5.3. Based on the results reported in the left most column of Figures 5.16, 5.17 and 5.18, it can be observed that, there is still a small impact of content dependency in the performance of the proposed method. This shows the opportunity of further improving the proposed method by adding content-dependent parameters in the feature set.

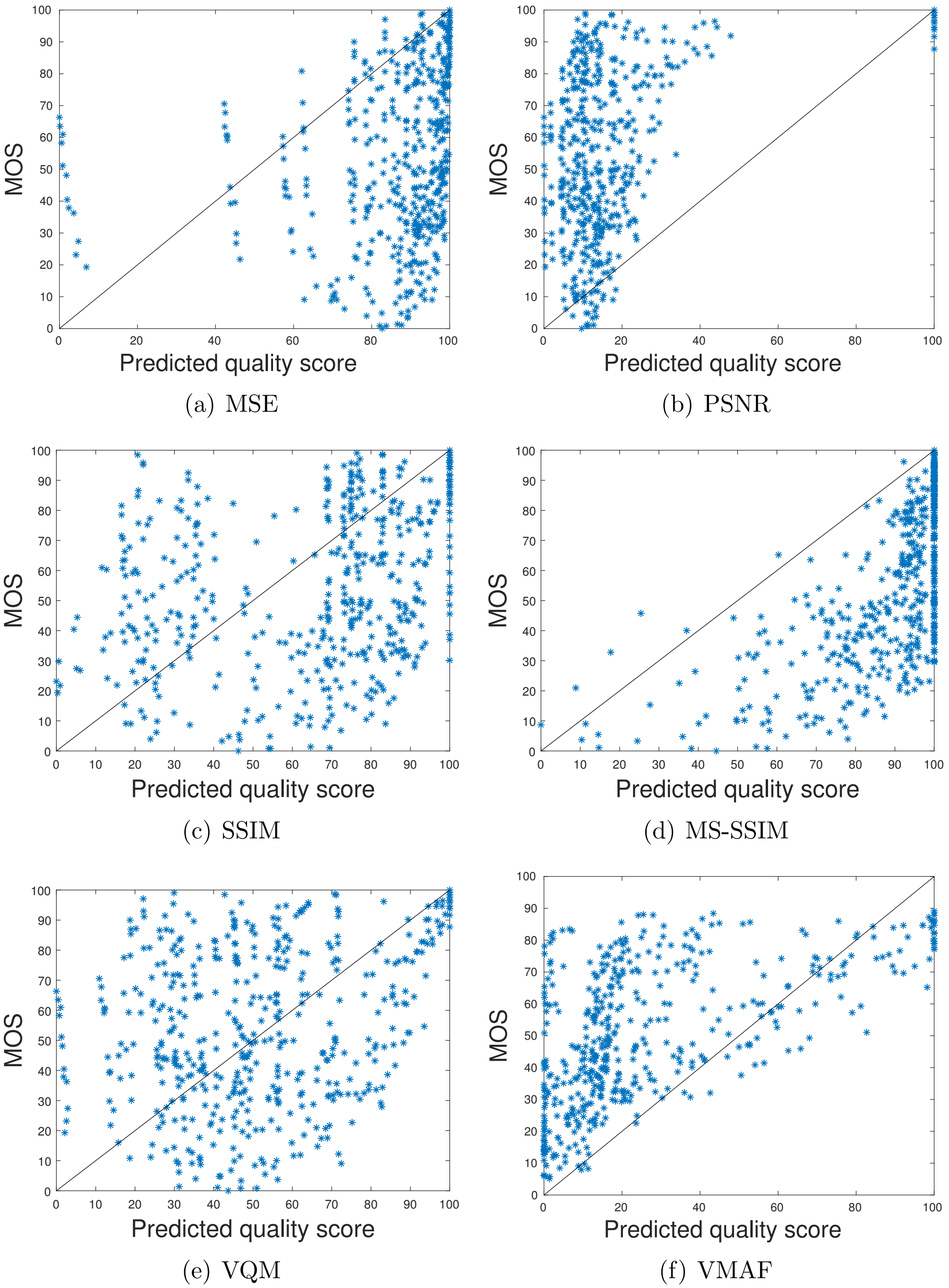
The proposed method used aliasing factors and MS-SSIM score as the inputs to training algorithm for predicting video quality. The spatio-temporal aliasing factors, as results in best performance, are designed to capture the quality degradation caused by frame rate and resolution change. On the other hand MS-SSIM is designed to predict quality degradation cause by compression. Therefore, it is expected that the combination of MS-SSIM scores and aliasing factors can cover degradations caused by the entire path of encoding as indicated in Fig 5.11.

We have also included the temporal motion smoothness (TMS) features proposed in Chapter 4 to train the overall VQA algorithm, and performed direct comparisons between the cases with and without using the TMS features. The results are given in Table 5.4.

When compared with Table 5.3, it suggests that adding the TMS factor into the feature set does not lead to further performance improvement. There might be two reasons. First, as has been shown in Chapter 4, the strength of TMS is highly content dependent, and thus in order to make use of TMS as a feature, certain content-dependent normalization process is needed. Second, there may be significant overlap or redundancy between the TMS and the aliasing factors, i.e., a video that is more smooth in motion is likely to create less aliasing. In such cases, adding TMS as an extra feature is not expected to enhance the quality prediction performance.

Table 5.4: The performance of the selected proposed methods with/without TMS factor on predicting the quality of videos in IVC-HFRVQA-II dataset.

Method	SRCC	PLCC	RMSE
Proposed A	0.7063	0.7212	27.95
Proposed B -T + TMS (SVR(MS-SSIM, TMS))	0.7252	0.7395	22.5902
Proposed B (SVR(MS-SSIM,T))	0.7352	0.7183	20.77
Proposed B + TMS (SVR(MS-SSIM, T, TMS))	0.7180	0.7298	22.1674
Proposed F (SVR(MS-SSIM,XY,T))	0.9592	0.9597	7.8139
Proposed F - T + TMS (SVR(MS-SSIM,XY,TMS))	0.9539	0.9543	8.0403



(a) MSE

(b) PSNR

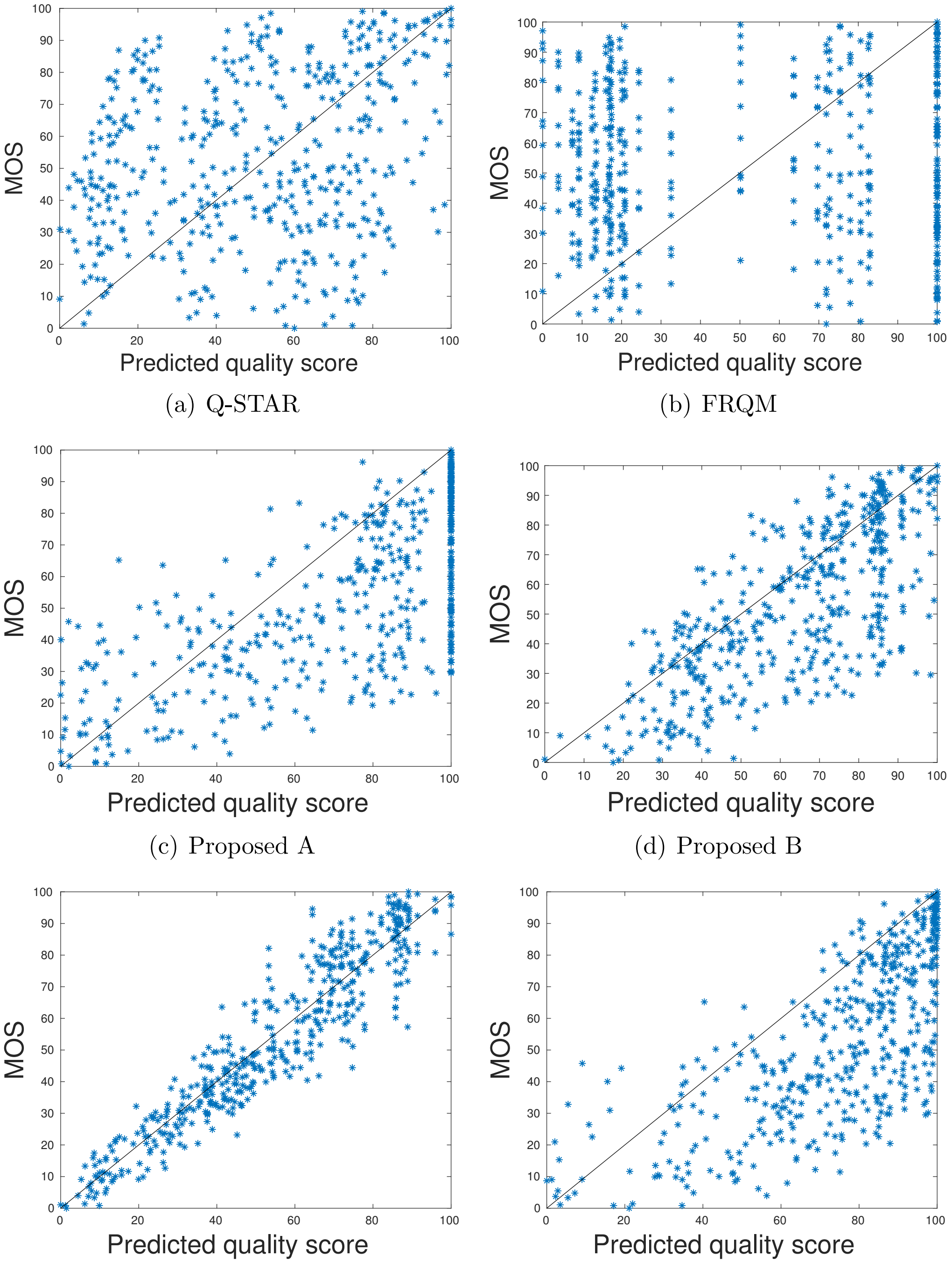
(c) SSIM

(d) MS-SSIM

(e) VQM

(f) VMAF

Figure 5.13: Scatter plots of the predicted values by VQA methods vs. MOS values on videos from IVC-HFRVQA-II dataset (cont'd).

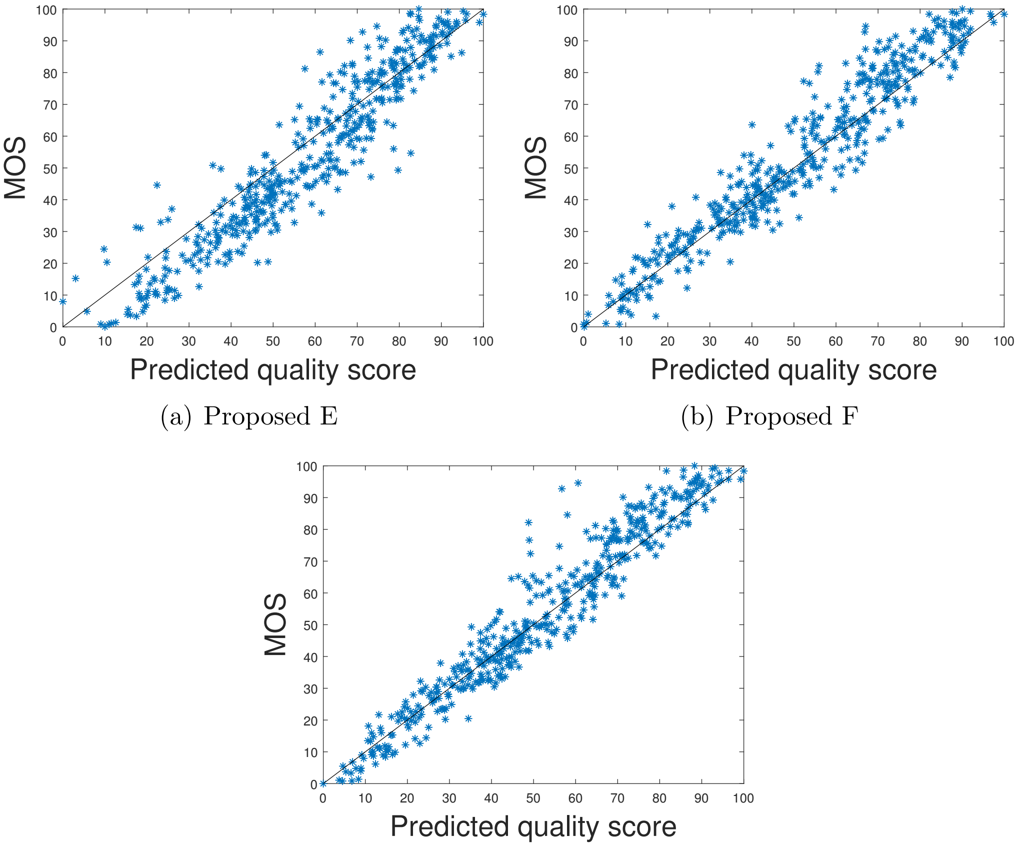


(a) Q-STAR (b) FRQM

(c) Proposed A (d) Proposed B

(e) Proposed C (f) Proposed D

Figure 5.14: Scatter plots of the predicted values by VQA methods vs. MOS values on videos from IVC-HFRVQA-II dataset (cont'd).



(a) Proposed E

(b) Proposed F

(c) Proposed G

Figure 5.15: Scatter plots of the predicted values by VQA methods vs. MOS values on videos from IVC-HFRVQA-II dataset(cont'd).

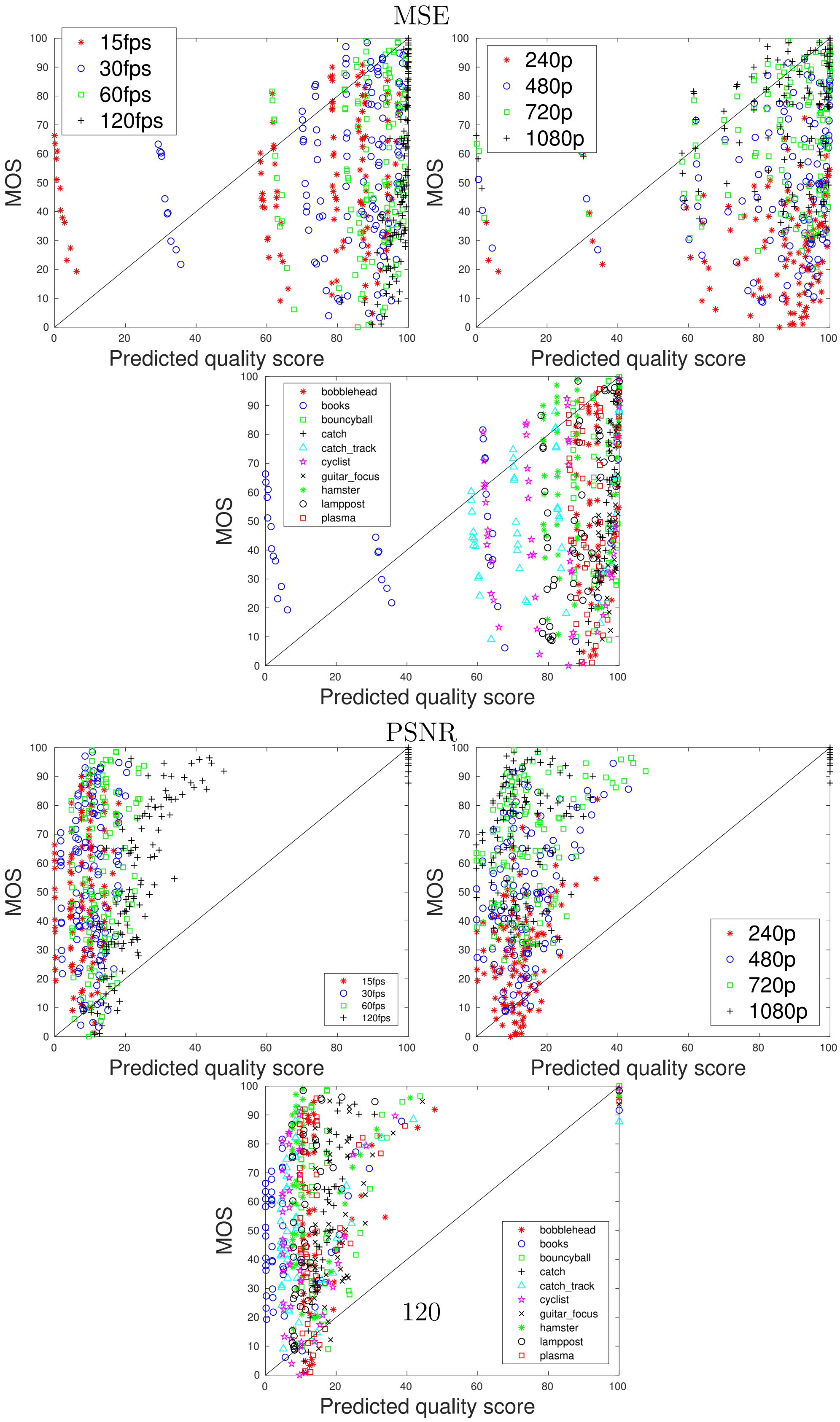


Figure 5.16: Scatter plots of the predicted values by VQA methods vs. MOS values on videos from IVC-HFRVQA-II dataset.

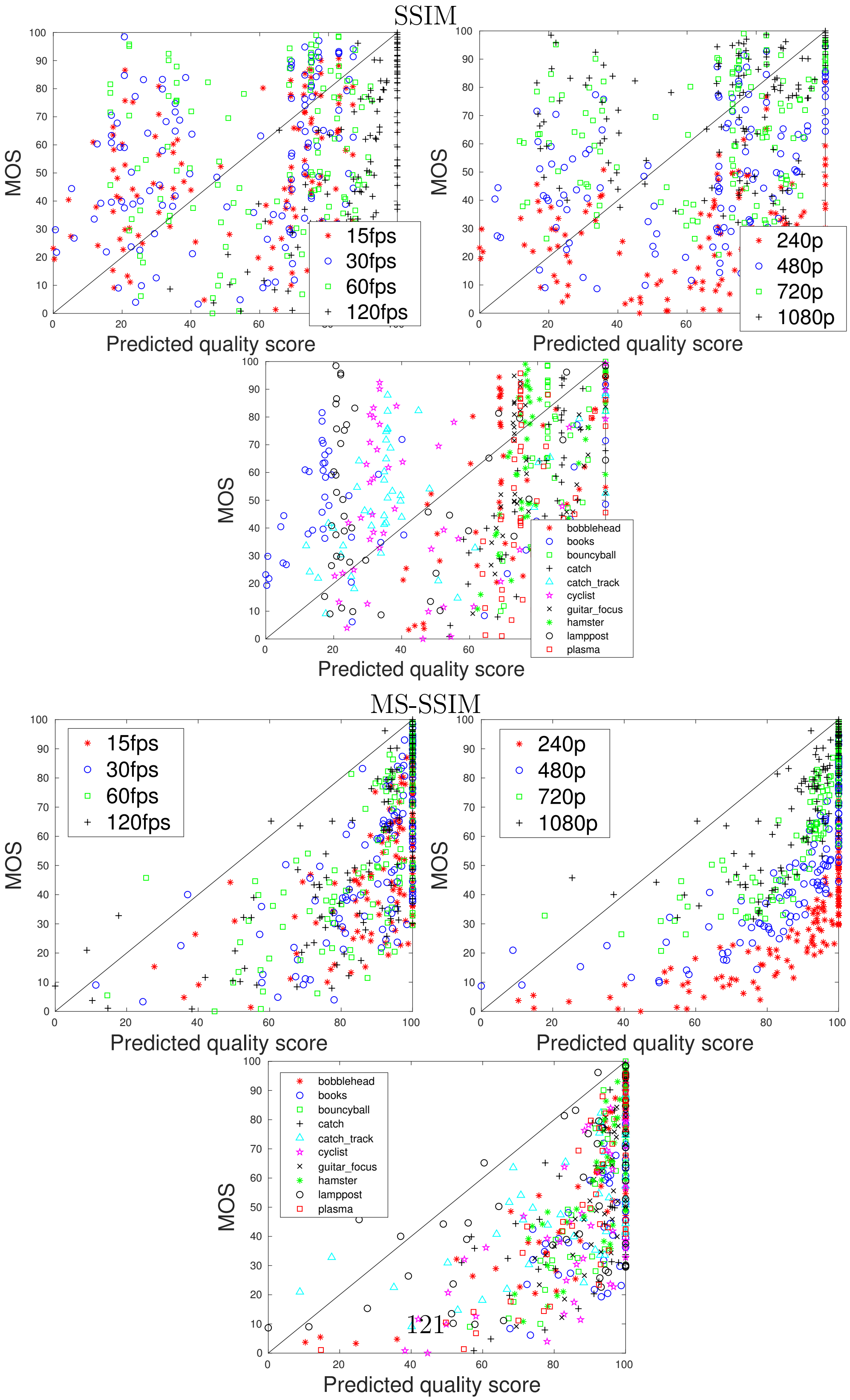
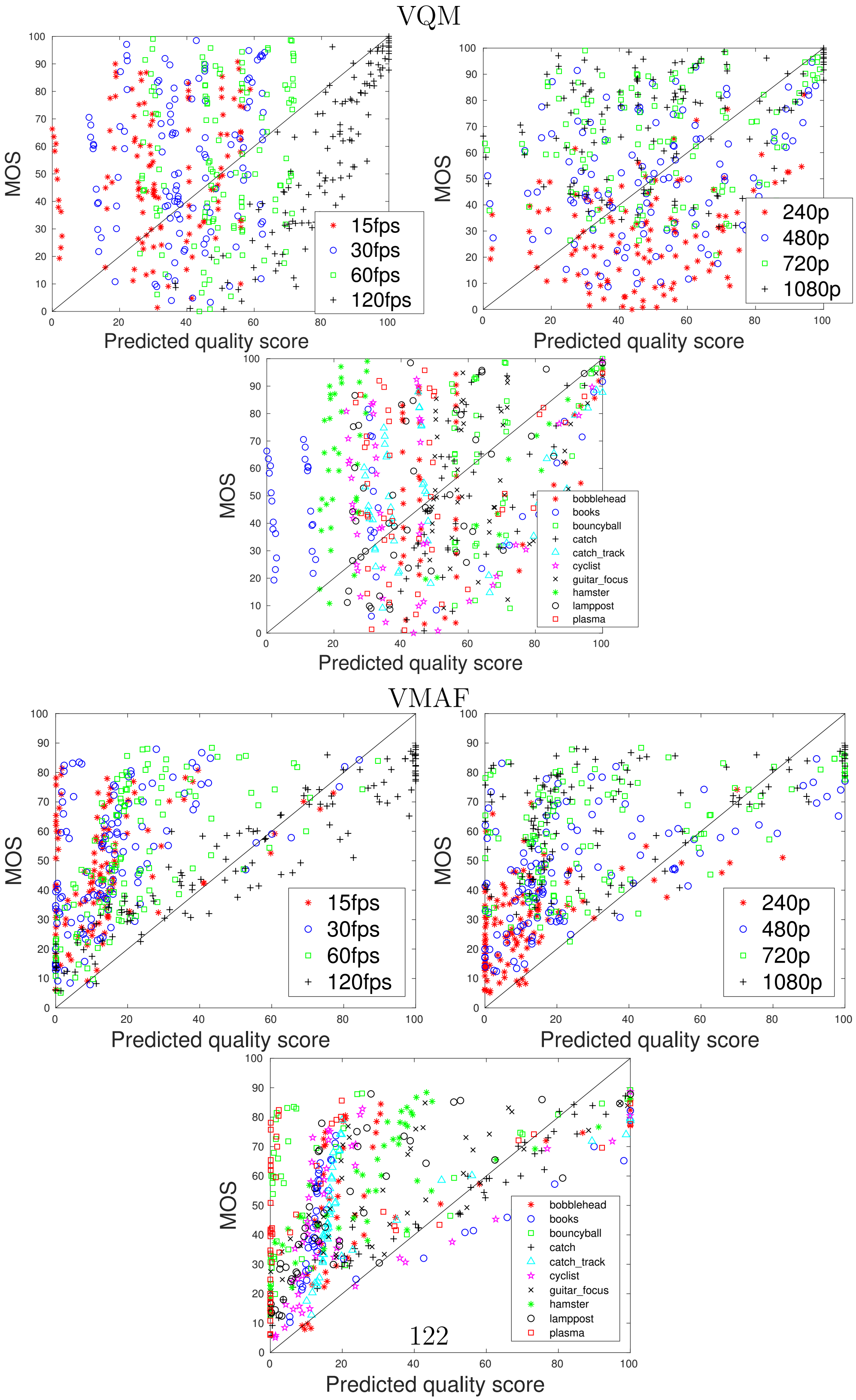


Figure 5.17: Scatter plots of the predicted values by VQA methods vs. MOS values on videos from IVC-HFRVQA-II dataset (cont'd).


VQM

VMAF

Figure 5.18: Scatter plots of the predicted values by VQA methods vs. MOS values on videos from IVC-HFRVQA-II dataset (cont'd).

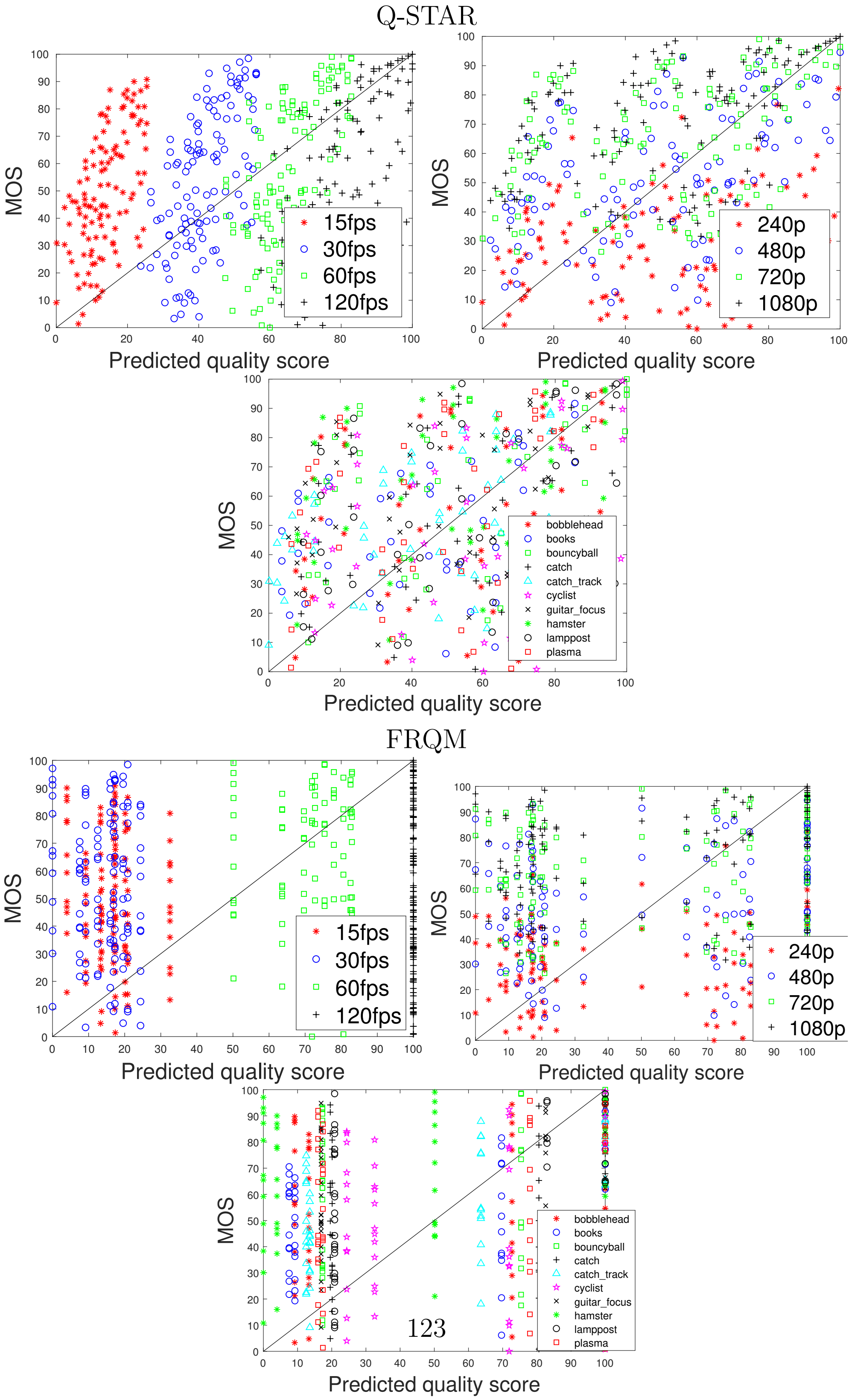


Figure 5.19: Scatter plots of the predicted values by VQA methods vs. MOS values on videos from IVC-HFRVQA-II dataset (cont'd).

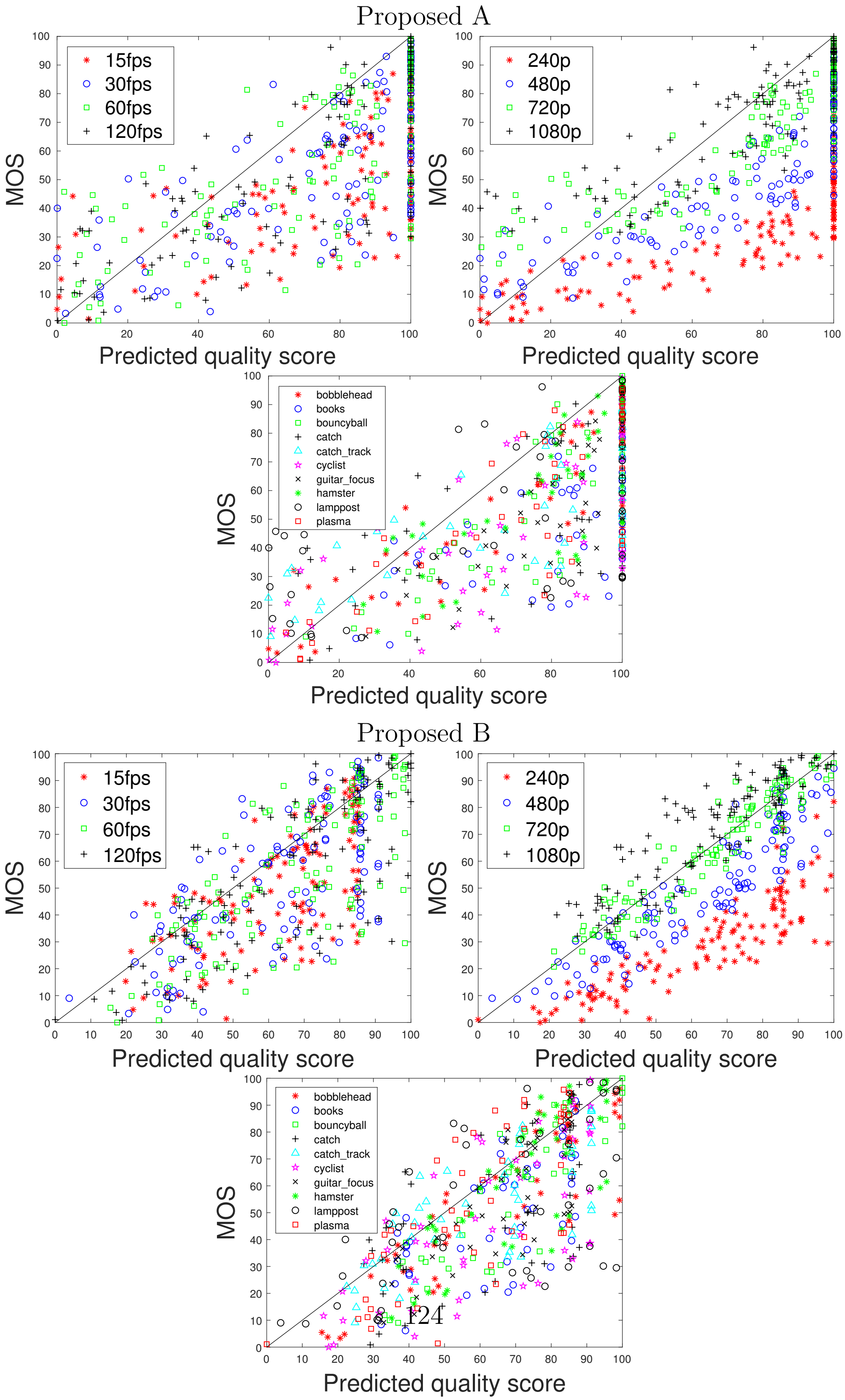
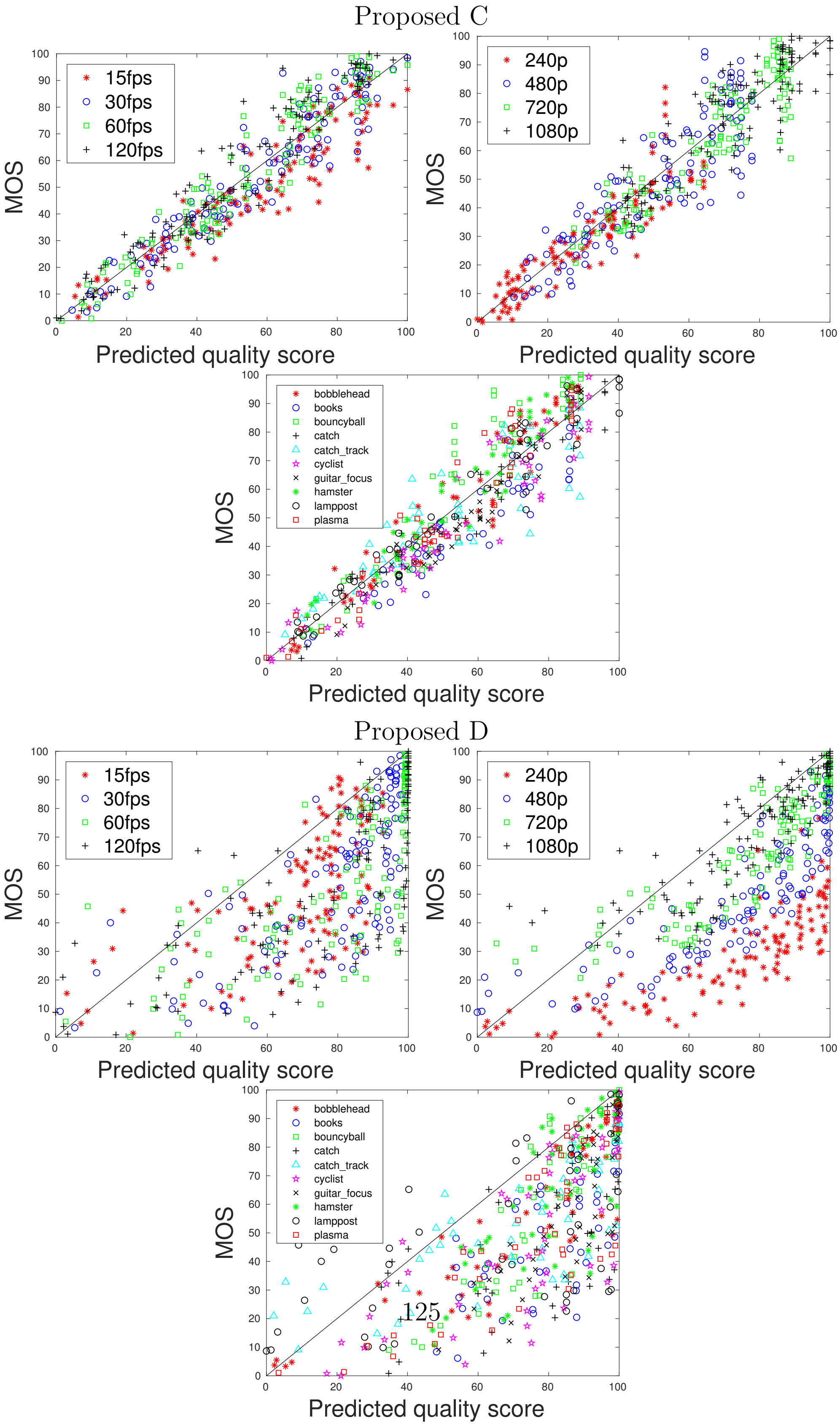


Figure 5.20: Scatter plots of the predicted values by VQA methods vs. MOS values on videos from IVC-HFRVQA-II dataset (cont'd).


Proposed C

Proposed D

Figure 5.21: Scatter plots of the predicted values by VQA methods vs. MOS values on videos from IVC-HFRVQA-II dataset (cont'd).

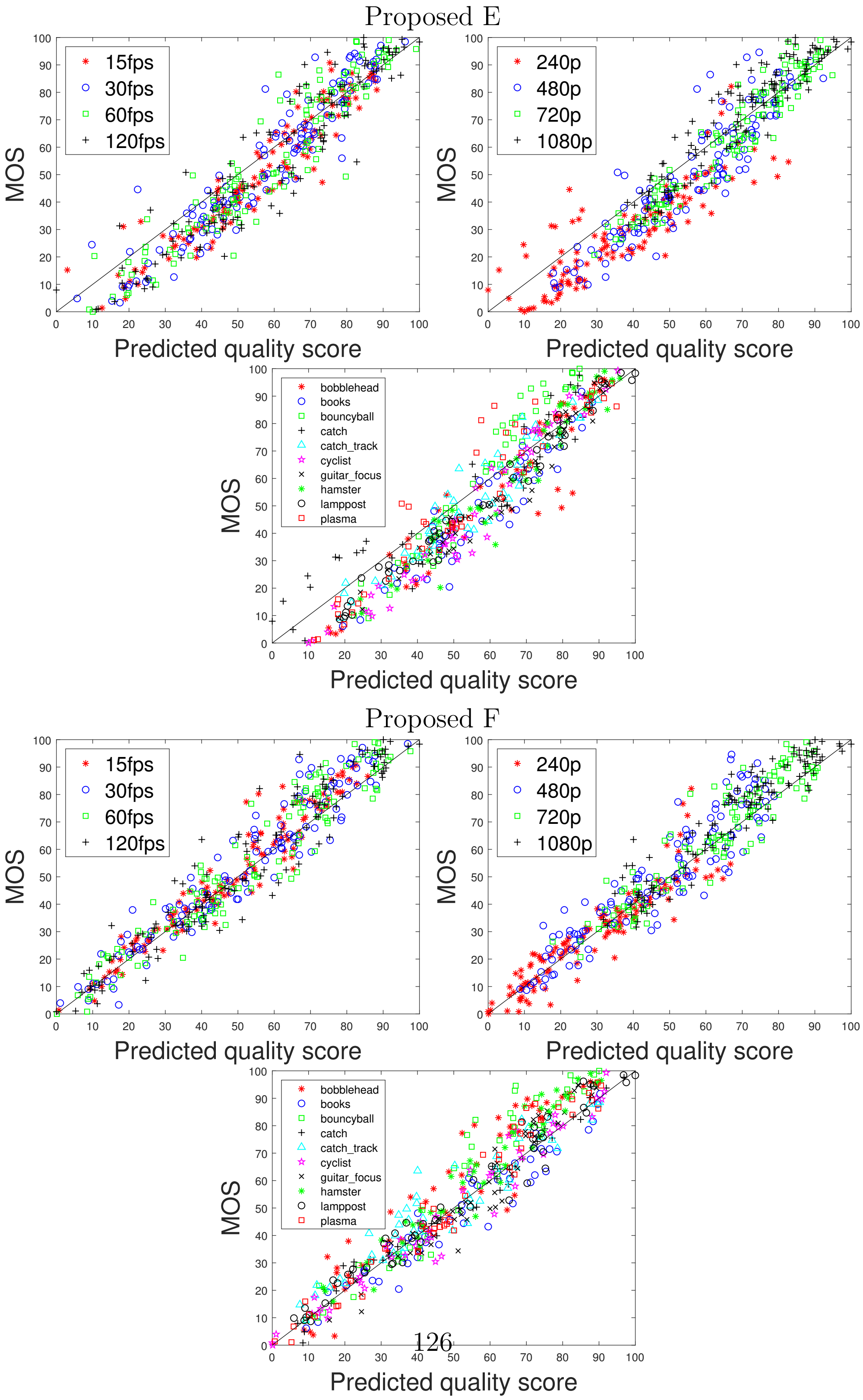


Figure 5.22: Scatter plots of the predicted values by VQA methods vs. MOS values on videos from IVC-HFRVQA-II dataset (cont'd).

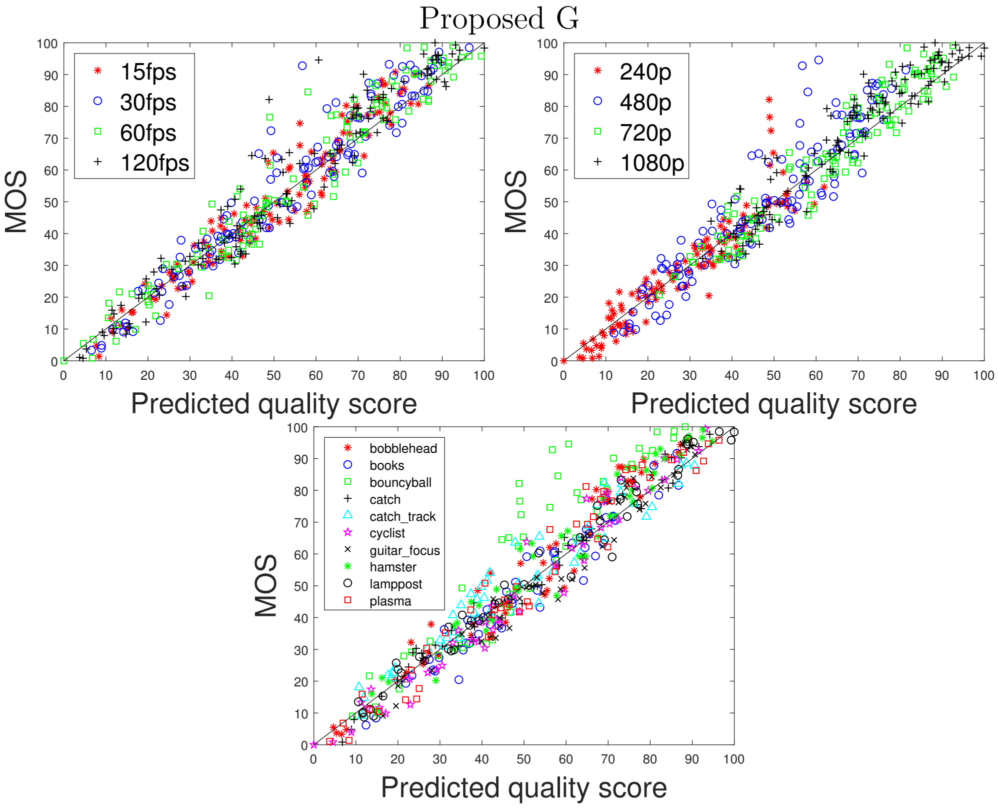


Figure 5.23: Scatter plots of the predicted values by VQA methods vs. MOS values on videos from IVC-HFRVQA-II dataset (cont'd).

5.3 Summary

In this chapter, we investigated the impact of frame rate changes from the perspective of human visual system models. Specifically, we examined the aliasing power of videos after frame rate and resolution reduction. We proposed the notion of perceptual aliasing factor, where the aliasing power is weighted by human visual contrast sensitivity along temporal direction or in spatio-temporal domain. The proposed perceptual temporal and spatio-temporal aliasing factors demonstrated good promise in cross-frame rate VQA. The proposed aliasing factor in combination with a frame rate-blind VQA model namely MS-SSIM creates a comprehensive model considering frame rate changes, resolution changes, and compression artifacts together. The performance analysis shows that it outperforms well-known VQA models on videos containing frame rate and resolution changes.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The main focus of this thesis is to understand the impact of frame rate on video quality, and to develop perceptual VQA models that produce consistent quality scoring on videos undergoing frame rate change, resolution change, and compression. To evaluate the subjective quality of video when the frame rate changes, we construct two video databases and perform subjective studies on these databases to investigate the quality of videos at different quantization levels, spatial resolutions and frame rates. Our overall observation is that, quality has direct relationship with frame rate, but the impact is more significant for frame rates lower than 60 fps. Depending on the content, the impact of resolution reduction on video quality could be dominant over the impact of frame rate reduction, especially at high frame rates. We also evaluate well-known VQA methods in predicting video quality using our databases. Our results show that the performance of existing VQA methods is limited in predicting human opinions.

The quality of videos with changing frame rate is investigated from the perspective of statistical analysis of motion in video. A statistical model on the temporal correlations of complex wavelet transform coefficient is constructed to measure temporal motion smoothness of videos. We found that temporal motion smoothness has direct relationship with the change in frame rate. It suggests that temporal motion smoothness may be a useful factor for VQA. On the other hand, content dependency is observed that has strong impact on the correlation of these two factors.

Significant effort has been dedicated to investigating the impact of frame rate changes from the perspective of human visual system models. We examined the aliasing power of videos after frame rate and resolution reduction and proposed the notion of perceptual aliasing factor, for which the aliasing power is weighted by human visual contrast sensitivity along temporal or spatio-temporal dimensions. The proposed aliasing factor in combination with a frame rate-blind VQA method namely MS-SSIM creates a comprehensive model that jointly considers frame rate changes, resolution changes, and compression artifacts together. Performance analysis shows that the joint model outperforms well-known VQA models when the test videos contain a mixture of quality degradations caused by frame rate changes, resolution changes, and compression.

6.2 Future Work

The current work may be extended in different ways.

Extension of Subjective Study: One meaningful further step in extending this work is to construct a bigger database of videos with more contents and performing a subjective study to gain a more reliable statistical analysis of the results. One common problem in subjective VQA studies is the trade-off between the number of the videos and the limited time for subjective testing in a session. The latter is strongly constrained by visual fatigue effect. Running another subjective test in addition to the current one with richer video content could help in this regard. Extending the experiment with crowd-sourcing is desirable but should be carried out with caution because of the complication in setting up a reliable and controllable visual testing environment.

Exploration of Content Dependent Features: One of the observations from the subjective study and the evaluation of VQA models in cross-frame rate video database is the strong content dependency, especially for frame rates higher than 30 fps. This content dependency needs to be investigated more deeply by exploiting content dependent features so as to better predict the impact of frame rate changes on different video content. This need is highlighted in the analysis of the proposed motion smoothness factor. Classifying the videos into different groups using content dependent features and designing or learning separate models for each group would be a potential direction. The new content dependent features could be from low-level signal analysis of video signal or higher-level concepts of motion and objects in the video. One reasonable way to find such features is to investigate temporal artifacts and related features to frame rate reduction. The appearance of video artifacts and the sensitivity of human visual perception are different to different types of video degradations.

Applications to Rate-Distortion Optimization: One extension of significant potential is to use the proposed VQA model for rate-distortion optimization, which is one of the most important applications of objective VQA models in video coding and transmission. Joint frame rate, spatial resolution, and quantization parameter optimization based on the proposed model can be investigated to achieve high-performance video coding using perceptual rate-distortion optimization to achieve the best perceptual quality at any desired bit rate. The modeling along the temporal direction may help video encoders to find the optimal trade-off point between bit rate and quality in a space that includes frame rate as a key parameter. This encoding and quality evaluation loop may be embedded in state-of-the-art video compression and streaming algorithms to improve coding and transmission efficiency in visual communication networks.

Bibliography

[1] K. Zeng, T. Zhao, A. Rehman, and Z. Wang, "Characterizing perceptual artifacts in compressed video streams," in *Proc. SPIE 9014, Human Vision and Electronic Imaging XIX*, San Francisco, CA, Feb. 2014.

[2] Rasoul Mohammadi Nasiri, Zhengfang Duanmu, and Zhou Wang, "Temporal motion smoothness and the impact of frame rate variation on video quality," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1418–1422.

[3] Rasoul Mohammadi Nasiri and Zhou Wang, "Perceptual aliasing factors and the impact of frame rate on video quality," in *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 3475–3479.

[4] DH Kelly, "Motion and vision. ii. stabilized spatio-temporal threshold surface," *Josa*, vol. 69, no. 10, pp. 1340–1349, 1979.

[5] Cisco Visual Networking Index, "Forecast and methodology, 2016-2021 white paper," *Retrieved 15rd October*, 2018.

[6] Cisco Visual Networking Index Cisco, "Global mobile data traffic forecast update, 2013–2018," *White paper*, 2014.

[7] Zhou Wang and Alan C Bovik, "Modern image quality assessment," *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, no. 1, pp. 1–156, 2006.

[8] Juan Pedro López Velasco, *Video quality assessment*, INTECH Open Access Publisher, 2012.

[9] Rajiv Soundararajan and Alan C Bovik, "Survey of information theory in visual quality assessment," *Signal, Image and Video Processing*, vol. 7, no. 3, pp. 391–401, 2013.

[10] Maria Torres Vega, Maria Torres Vega, Vittorio Sguazzo, Vittorio Sguazzo, Decebal Constantin Mocanu, Decebal Constantin Mocanu, Antonio Liotta, and Antonio Liotta, "An experimental survey of no-reference video quality assessment methods," *International Journal of Pervasive Computing and Communications*, vol. 12, no. 1, pp. 66–86, 2016.

[11] Paul Read and Mark-Paul Meyer, *Restoration of motion picture film*, Butterworth-Heinemann, 2000.

[12] Damon M Chandler, "Seven challenges in image quality assessment: past, present, and future research," *ISRN Signal Processing*, vol. 2013, 2013.

[13] P ITU-T RECOMMENDATION, "910," *Subjective video quality assessment methods for multimedia applications*, pp. 910–200804, 2008.

[14] P. Brooks and B. Hestnes, "User measures of quality of experience: why being objective and quantitative is important," *IEEE Network*, vol. 24, no. 2, pp. 8–13, March 2010.

[15] P ITU-T RECOMMENDATION, "Subjective video quality assessment methods for multimedia applications," Apr 2008.

[16] Zhou Wang and Alan 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.

[17] Quan Huynh-Thu and Mohammed Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.

[18] Hui Li Tan, Zhengguo Li, Yih Han Tan, Susanto Rahardja, and Chuohuo Yeo, "A perceptually relevant mse-based image quality metric," *IEEE Transactions on Image Processing*, vol. 22, no. 11, pp. 4447–4459, 2013.

[19] XK Yang, WS Lin, Zhongkang Lu, Ee Ping Ong, and Susu Yao, "Just-noticeable-distortion profile with nonlinear additivity model for perceptual masking in color images," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. IEEE, 2003, vol. 3, pp. III–609.

[20] Xinbo Gao, Wen Lu, Dacheng Tao, and Xuelong Li, "Image quality assessment based on multiscale geometric analysis," *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1409–1423, 2009.

[21] Zhou Wang, Alan C Bovik, and Ligang Lu, "Why is image quality assessment so difficult?," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 4, pp. IV–3313.

[22] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[23] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference*

Record of the Thirty-Seventh Asilomar Conference on. Ieee, 2003, vol. 2, pp. 1398–1402.

[24] Zhou Wang and Qiang Li, “Information content weighting for perceptual image quality assessment,” *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2011.

[25] Christian Szegedy, Alexander Toshev, and Dumitru Erhan, “Deep neural networks for object detection,” in *Advances in neural information processing systems*, 2013, pp. 2553–2561.

[26] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov, “Scalable object detection using deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2147–2154.

[27] Sebastian Bosse, Dominique Maniry, Thomas Wiegand, and Wojciech Samek, “A deep neural network for image quality assessment,” in *2016 IEEE International Conference on Image Processing (ICIP).* IEEE, 2016, pp. 3773–3777.

[28] Le Kang, Peng Ye, Yi Li, and David Doermann, “Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks,” in *2015 IEEE international conference on image processing (ICIP).* IEEE, 2015, pp. 2791–2795.

[29] Weilong Hou, Xinbo Gao, Dacheng Tao, and Xuelong Li, “Blind image quality assessment via deep learning,” *IEEE transactions on neural networks and learning systems*, vol. 26, no. 6, pp. 1275–1286, 2015.

[30] Jongyoo Kim and Sanghoon Lee, “Fully deep blind image quality predictor,” *IEEE Journal of selected topics in signal processing*, vol. 11, no. 1, pp. 206–220, 2017.

[31] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2018.

[32] Jongyoo Kim, Hui Zeng, Deepti Ghadiyaram, Sanghoon Lee, Lei Zhang, and Alan C Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 130–141, 2017.

[33] Le Kang, Peng Ye, Yi Li, and David Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1733–1740.

[34] Jie Xiang Yang and Hong Ren Wu, "Robust filtering technique for reduction of temporal fluctuation in h. 264 video sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 3, pp. 458–462, 2010.

[35] BT ITU-R RECOMMENDATION, "Methodology for the subjective assessment of video quality in multimedia applications," Jan 2007.

[36] BT ITU-R RECOMMENDATION, "Mmethodology for the subjective assessment of the quality of television pictures," Jan 2012.

[37] VQEG Official Web Site, "Subjective assessment methods for image quality in high-definition television," *Rec. ITU-R BT. 710-4*.

[38] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack, "Study of subjective and objective quality assessment of video," *IEEE transactions on image processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[39] Zhou Wang, Ligang Lu, and Alan C Bovik, “Video quality assessment based on structural distortion measurement,” *Signal processing: Image communication*, vol. 19, no. 2, pp. 121–132, 2004.

[40] Zhou Wang and Qiang Li, “Video quality assessment using a statistical model of human visual speed perception,” *JOSA A*, vol. 24, no. 12, pp. B61–B69, 2007.

[41] Alan A Stocker and Eero P Simoncelli, “Noise characteristics and prior expectations in human visual speed perception,” *Nature neuroscience*, vol. 9, no. 4, pp. 578–585, 2006.

[42] Margaret H Pinson and Stephen Wolf, “A new standardized method for objectively measuring video quality,” *IEEE Transactions on broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[43] Kalpana Seshadrinathan and Alan C Bovik, “Motion-based perceptual quality assessment of video,” in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2009, pp. 72400X–72400X.

[44] Kalpana Seshadrinathan and Alan Conrad Bovik, “Motion tuned spatio-temporal quality assessment of natural videos,” *IEEE transactions on image processing*, vol. 19, no. 2, pp. 335–350, 2010.

[45] J. Y. C. Chen and J. E. Thropp, “Review of low frame rate effects on human performance,” *IEEE Trans. System, Man and Cybernetics, Part A: Systems and Humans*, vol. 37, no. 6, pp. 1063–1076, Nov. 2007.

[46] Z. Wang and Q. Li, “Video quality assessment using a statistical model of human visual speed perception,” *Journal of the Optical Society of America*, vol. 24, no. 12, pp. B61–B69, Dec. 2007.

[47] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, "Cross-dimensional perceptual quality assessment for low bit-rate videos," *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1316–1324, Nov. 2008.

[48] Y. Ou, T. Liu, Z. Zhao, Z. Ma, and Y. Wang, "Modeling the impact of frame rate on perceptual quality of video," in *Proc. IEEE Int. Conf. Image Proc.*, San Diego, CA, Oct 2008, pp. 689–692.

[49] Y. Ou, Z. Ma, and Y. Wang, "Modeling the impact of frame rate and quantization stepsizes and their temporal variations on perceptual video quality: A review of recent works," in *Proc. IEEE Int. Conf. on Information Sciences and Systems*, Princeton, NJ, March 2010, pp. 1–6.

[50] Y. Ou, Z. Ma, T. Liu, and Y. Wang, "Perceptual quality assessment of video considering both frame rate and quantization artifacts," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 21, no. 3, pp. 286–298, 2011.

[51] L. Janowski and P. Romaniak, "QoE as a function of frame rate and resolution changes," in *Future Multimedia Networking*, pp. 34–45. 2010.

[52] Jose Joskowicz and J Ardao, "Combining the effects of frame rate, bit rate, display size and video content in a parametric video quality model," in *Proceedings of the 6th Latin America Networking Conference.* ACM, 2011, pp. 4–11.

[53] Q. Huynh-Thu and M. Ghanbari, "Temporal aspect of perceived quality in mobile video broadcasting," *IEEE Trans. on Broadcasting*, vol. 54, no. 3, pp. 641–651, Sep. 2008.

[54] Ming-Chen Chien, Ren-Jie Wang, Chien-Hsun Chiu, and Pao-Chi Chang, "Quality driven frame rate optimization for rate constrained video encoding," *Broadcasting, IEEE Transactions on*, vol. 58, no. 2, pp. 200–208, 2012.

[55] Gayatri Yadavalli, Mark Masry, and Sheila S Hemami, "Frame rate preferences in low bit rate video," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*. IEEE, 2003, vol. 1, pp. I–441.

[56] J. Joskowicz, R. Sotelo, and Ardao J. C. L., "Towards a general parametric model for perceptual video quality estimation," *IEEE Trans. on Broadcasting*, vol. 59, no. 4, pp. 569–579, Dec. 2013.

[57] A. Khan, L. Sun, and E. Ifeachor, "QoE prediction model and its application in video quality adaptation over umts networks," *IEEE Trans. on Multimedia*, vol. 14, no. 2, pp. 431–442, Apr. 2012.

[58] M. Claypool, K. Claypool, and F. Damaa, "The effects of frame rate and resolution on users playing first person shooter games," in *Proc. SPIE 6071, Multimedia Computing and Networking*, San Jose, CA, Jan. 2006.

[59] A. Banitalebi-Dehkordi, M. T. Pourazad, and P. Nasiopoulos, "Effect of high frame rates on 3D video quality of experience," in *Proc. IEEE Int. Conf. on Cons. Electron.*, Las Vegas, NV, Jan. 2014, pp. 416–417.

[60] A. Banitalebi-Dehkordi, M. T. Pourazad, and P. Nasiopoulos, "The effect of frame rate on 3D video quality and bitrate," *3D Research*, vol. 6, no. 1, pp. 1–13, Dec. 2014.

[61] Fan Zhang, Alex Mackin, and David R Bull, "A frame rate dependent video quality metric based on temporal wavelet decomposition and spatiotemporal pooling," in

2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017, pp. 300–304.

[62] Z Li, A Norkin, and A Aaron, "Vmaf-video quality metric alternative to psnr," *Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11*, 2016.

[63] Christos G Bampis, Zhi Li, and Alan C Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[64] Christos G Bampis and Alan C Bovik, "Learning to predict streaming video qoe: Distortions, rebuffering and memory," *arXiv preprint arXiv:1703.00633*, 2017.

[65] Reza Rassool, "Vmaf reproducibility: Validating a perceptual practical video quality metric," in *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 2017, pp. 1–2.

[66] Chulhee Lee, S Woo, S Baek, J Han, J Chae, and J Rim, "Comparison of objective quality models for adaptive bit-streaming services," in *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*. IEEE, 2017, pp. 1–4.

[67] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication, special issue on Objective video quality metrics*, vol. 19, no. 2, pp. 121–132, Feb. 2004.

[68] Alex Mackin, Fan Zhang, and David R Bull, "A study of subjective video quality at various frame rates," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 3407–3411.

[69] Rasoul Mohammadi Nasiri, Jiheng Wang, Abdul Rehman, Shiqi Wang, and Zhou Wang, "Perceptual quality assessment of high frame rate video," in *Multimedia Signal Processing (MMSP), 2015 IEEE 17th International Workshop on*. IEEE, 2015, pp. 1–6.

[70] RECOMMENDATION ITU-R BT, "Methodology for the subjective assessment of the quality of television pictures," 2009.

[71] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack, "Study of subjective and objective quality assessment of video," *IEEE transactions on image processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[72] Stephen Wolf and M Pinson, "Application of the ntia general video quality metric (vqm) to hdtv quality monitoring," in *Proceedings of The Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Scottsdale, AZ, USA*, 2007.

[73] Jinggang Huang and David Mumford, "Statistics of natural images and models," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on*. IEEE, 1999, vol. 1.

[74] Christopher Kanan and Garrison Cottrell, "Robust classification of objects, faces, and flowers using natural image statistics," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2472–2479.

[75] Bruno A Olshausen and David J Field, "Natural image statistics and efficient coding," *Network: computation in neural systems*, vol. 7, no. 2, pp. 333–339, 1996.

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

[77] Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on image processing*, vol. 14, no. 12, pp. 2117–2128, 2005.

[78] Zhou Wang and Eero P Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Electronic Imaging 2005*. International Society for Optics and Photonics, 2005, pp. 149–159.

[79] Anush Krishna Moorthy and Alan Conrad Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[80] Zhou Wang and Qiang Li, "Statistics of natural image sequences: temporal motion smoothness by local phase correlations," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2009, pp. 72400W–72400W.

[81] Kai Zeng and Zhou Wang, "Temporal motion smoothness measurement for reduced-reference video quality assessment.," in *ICASSP*, 2010, pp. 1010–1013.

[82] Nicholas I Fisher, *Statistical analysis of circular data*, Cambridge University Press, 1995.

[83] Kantilal Varichand Mardia, *Statistics of directional data*, Academic press, 2014.

[84] Scott J Daly, "Engineering observations from spatiovelocity and spatiotemporal visual models," in *Photonics West'98 Electronic Imaging*. International Society for Optics and Photonics, 1998, pp. 180–191.

[85] A Murat Demirtas, Amy R Reibman, and Hamid Jafarkhani, "Full reference video quality estimation for videos with different spatial resolutions," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 1997–2001.

[86] "MS-SSIM: Multi-scale structural similarity for image quality assessment," https://ece.uwaterloo.ca/~z70wang/research/iwssim, Accessed: 2019-01-10.

[87] "VQM: Video quality metric (vqm) software," https://www.its.bldrdoc.gov/resources/video-quality-research/software.aspx, Accessed: 2019-1-10.

[88] Yen-Fu Ou, Yuanyi Xue, and Yao 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.

[89] "VMAF: Video multi-method assessment fusion," https://github.com/Netflix/vmaf, Accessed: 2019-1-10.