UNIVERSITY OF CALIFORNIA,
IRVINE

Quality Estimation and Delivery of Visual Contents with Different Spatial Resolutions

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Electrical and Computer Engineering

by

A. Murat Demirtas

Dissertation Committee:
Professor Hamid Jafarkhani, Chair
Professor Ender Ayanoglu
Associate Professor Athina Markopoulou

2015

DEDICATION

To my parents and my brother

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to take this opportunity to thank the people who have stood by me and have helped me accomplish this work. First, I would like to thank my advisor, Prof. Hamid Jafarkhani, for his constant support and guidance. He is an excellent teacher and researcher and I thank him for all our discussions which shape my research. In spite of his busy schedule, he was very patient and always had time to meet with me.

I would like to thank my unofficial advisor, Prof. Amy Reibman from Purdue University for being an excellent mentor in guiding my research. The credit for my research topic goes to her, who introduced me to this topic while I was interning at AT&T Labs. I would like to specially thank for her detailed comments in our weekly meetings which makes our long distance collaboration a success.

I would like to my committee members, Professor Ender Ayanoglu and Professor Athina Markopolou for their invaluable time and constructive comments.

I would also like to thank my labmates for their constructive comments to my research.

Last but not least, I would like to thank my family for their unconditional love and support. None of my achievements throughout my life would have been possible without their encouragement.

CURRICULUM VITAE

A. Murat Demirtas

EDUCATION

Doctor of Philosophy in Electrical and Computer Engineering **2015**
University of California, Irvine *Irvine, CA*

Master of Science in Electrical and Electronics Engineering **2008**
Middle East Technical University *Ankara, Turkey*

Bachelor of Science in Electrical and Electronics Engineering **2005**
Bilkent University *Ankara, Turkey*

RESEARCH EXPERIENCE

Graduate Research Assistant **2011–2015**
University of California, Irvine *Irvine, CA*

Graduate Research Assistant **2009–2010**
Middle East Technical University, Ankara *Ankara, Turkey*

TEACHING EXPERIENCE

Teaching Assistant **2011–2015**
University of California, Irvine *Irvine, CA*

REFEREED JOURNAL PUBLICATIONS

Full Reference Video Quality Estimation for Videos with Different Spatial Resolutions IEEE Transactions on Circuits and Systems for Video Technology	**2015**
Full Reference Quality Estimation for Images with Different Spatial Resolutions IEEE Transactions on Image Processing	**2014**

REFEREED CONFERENCE PUBLICATIONS

Full Reference Video Quality Estimation for Videos with Different Spatial Resolutions International Conference on Image Processing (ICIP)	**Oct 2014**
Image Quality Estimation for Different Spatial Resolutions International Conference on Image Processing (ICIP)	**Sep 2013**
Quality Estimation for Images and Video with Different Spatial Resolutions Human Vision and Electronic Imaging (HVEI)	**Feb 2012**
Performance of H.264 with Isolated Bit Error: Packet Decode or Discard? International Conference on Image Processing (ICIP)	**Sep 2011**

ABSTRACT OF THE DISSERTATION

Quality Estimation and Delivery of Visual Contents with Different Spatial Resolutions

By

A. Murat Demirtas

Doctor of Philosophy in Electrical and Computer Engineering

University of California, Irvine, 2015

Professor Hamid Jafarkhani, Chair

In this thesis, we work on quality estimation and delivery of visual content with different spatial resolutions. First, we study the quality estimation of images with different spatial resolutions. Estimating the quality of the visual content accurately is crucial in providing satisfactory multimedia communication. State of the art visual quality assessment approaches are effective when the input image and the reference image have the same resolution. However, finding the quality of an image that has spatial resolution different than that of the reference image is still a challenging problem. To solve this problem, we develop a quality estimator (QE) which computes the quality of the input image without resampling the reference or the input images. We begin by identifying the potential weaknesses of previous approaches used to estimate the quality of experience. Next, we design a QE, called Multiscale Image Quality Estimator (MIQE), to estimate the quality of a distorted image with a lower resolution compared to the reference image. We also propose a subjective test environment to explore the success of the proposed algorithm in comparison with other QEs. When the input and test images have different resolutions, the subjective tests demonstrate that in most cases the proposed method works better than other approaches. In addition, the proposed algorithm also performs well when the reference image and the test image have the same resolution.

Second, we examine the quality estimation of videos with different spatial resolutions. Full-reference video QEs either resize the distorted input video or the reference video to compute the quality when these videos have different spatial resolutions. This resizing operation causes several limitations. MIQE overcomes those limitations for images but it does not consider the temporal characteristics of video. We develop a full-reference video quality estimator that integrates MIQE with the motion information to estimate the quality of the distorted video without resampling the reference or the test videos. We also perform subjective tests to compare the proposed algorithm with the existing QEs. In these tests, the reference and the input videos are displayed at their native resolutions. Test results show that the proposed algorithm outperforms other QEs when the reference video and the input video have different spatial resolutions. We have also evaluated the performance of the approach using the Scalable Video Database.

Third, we work on the challenge of using a perceptual quality estimator to perform optimum multicasting of videos to the devices with different spatial resolutions. We specifically focus on the complexity of the optimum perceptual multicasting. The complexity increases due to usage of scalable video coding in combined scalability mode and perceptual quality estimators. Using combined scalability increases the number of scalability options, so we need to perform multi-criteria optimization. As a result of the simulations we have performed, we have observed that multi-criteria optimization is not necessary in the low bitrate region, and we propose an algorithm to reduce the complexity of the optimization notably for this region.

Chapter 1

Introduction

1.1 Motivation and Previous Works

Multimedia communications have become ubiquitous as a result of progress in consumer products and social media abilities. Now, we can stream and watch any program we want from the Internet using Amazon, Netflix, Hulu. Even TV channels have started streaming their programs. We can also visually communicate with somebody instantly without taking long trips using programs like Skype and Gtalk. Furthermore, it is also possible to share video clips via YouTube. Hence, people have become the broadcasters. This trend has increased the wireless multimedia traffic substantially. According to estimations by Cisco [1], video traffic will be at least 69% of all internet traffic by 2017. Since the demand is increasing exponentially, the available wireless communication infrastructure may not be sufficient to provide a satisfactory service in the future. Therefore, we should develop methods to employ the resources efficiently. One of these methods is using perceptual coding approaches to minimize the required bitrate without causing any noticeable artifact. To achieve this, the quality of the visual content must be estimated with high accuracy.

Quality estimators (QEs) are used to measure the perceived quality of the visual content. Having an accurate QE is helpful to define system requirements, provide a benchmark to compare among different image and video processing algorithms, and solve a rate-distortion optimization problem related to visual communications or storage [2]. Objective QEs are evaluated in three categories: Full-Reference (FR) QE, Reduced-Reference (RR) QE and No-Reference (NR) QE. FR and RR QE approaches exploit the reference video as a whole and partially to estimate the quality of the test video, respectively. On the other hand, NR QE is independent of the reference video. To decide the type of the QE and design it accurately, we must take into account the requirements of the environment.

The display resolution of the end user is one of the important requirements to develop an accurate QE if the network consists of devices with various spatial resolution constraints. We frequently use devices like smart phones, laptops and HDTVs to perform various visual tasks. Viewing distance also affects the perceived spatial resolution. These features of viewing experience can change our perception of visual quality. Hence, the QE design should include the effect of resolution on the perception of quality. Having such a QE provides a solution to the utility issue in [3] for spatially adapted videos. Adaptive streaming services like MPEG-DASH can utilize this QE to provide high-quality spatially-scalable streaming. Quantifying the quality of visual sources at different resolutions can also be useful for designing displays for next generation of consumer products with very high resolution.

Image and video quality researchers have explored the effect of mixed spatio-temporal resolution change on the perceived quality of the visual content, both subjectively and objectively. The first group of studies only uses subjective tests to understand the influence of spatial or temporal resolution. Bae et al. [4] explore the preferred spatial resolution for a given level of quantization error. The authors show that viewers prefer to see larger images with a small amount of visible quantization error rather than a smaller image which has no visible quantization errors. They also demonstrate that the amount of acceptable distortion increases as the spatial resolution decreases. In [5] and [6], the authors perform subjective tests to examine the influence of jointly adjusting spatial resolution, temporal resolution, and quantization step-size on the quality. According to tests in [5], people prefer a smaller frame with smaller quantization errors compared to a larger frame with larger quantization errors, for the same bit-rate when the bitrate is low. On the other hand, Lee et al. [6] show that when the bitrate is high, increasing the temporal resolution is the most preferred option. Cermak et al. [7] use the mean opinion scores (MOS) to find the required bitrate to achieve a given level of video quality for a particular screen resolution. According to McCarthy et al. [8], people prefer high-quality video frames to smoothly moving video. Cranley et al. [9] perform subjective tests using several combinations of resolution and quantization parameters and formed an optimum path, called Optimum adaptation trajectory, to decide which parameters to change. Zhai et al. [10] examine the trade-off between parameters for low bitrates. They observe that enhancing the quantization level provides higher subjective quality compare to increasing spatio-temporal resolution of the video. These studies provide valuable information to understand the subjective effect of spatial resolution, but they do not provide an objective metric that can quantify this effect.

The second group of studies focuses on developing objective quality estimators and measuring their performance. We divide these studies into three subgroups: NR QEs, FR QEs that were not specifically designed to account for different resolutions, and FR QEs that were. An NR-QE does not use a reference image or video to estimate the quality. Instead, it may use the characteristics of possible artifacts like blocking or noise to estimate the quality, or it may use a model of the desired signal to identify deviations in the visual content from the model [11]. NR QE algorithms in [12, 13, 14, 15, 16, 17] have satisfactory results if the distorted visual content is not obtained by resampling. However, an NR QE cannot evaluate the degree of detail loss if the distorted input image/video has a lower resolution compare to the reference image/video. Hence, an NR QE will sacrifice valuable information if a reference is available.

The QE studies in the second subgroup have been designed to estimate quality when the reference and input images/videos have the same spatial resolutions. There are two straightforward methods to adapt these QEs to estimate the quality when the displayed content has a different spatial resolution. The first method compares the low-resolution input image/video with the decimated reference image/video. The second method compares the interpolated input image/video with the reference image/video. Throughout this document, we call these methods QE$_{down}$ and QE$_{up}$, respectively. The QEs in this subgroup use either the QE$_{down}$ or QE$_{up}$ strategy to compute quality for different resolutions. Thus, these QEs do not explicitly take into account the effect of spatial resolution. PSNR, Structural Similarity Index (SSIM) [18], Multiscale SSIM (MSSIM) [19], Video Quality Metric (VQM) [20] and Motion-based Video Integrity Evaluation (MOVIE) [21] are members of this subgroup. PSNR, SSIM and MSSIM are image QEs, but we can apply them to video by averaging the quality of the frames. On the other hand, VQM and MOVIE are video QEs.

Both QE$_{down}$ and QE$_{up}$ have three significant drawbacks. The first drawback is the need to resize the input or the reference visual content to estimate the quality. In QE$_{down}$, high-frequency content is lost from the reference image/video prior to comparison. On the other hand, in QE$_{up}$, the input image/video is further degraded before the comparison due to the upsampling filter. Hence, the resizing operation causes over- or under-estimation of the input visual content's quality in QE$_{down}$ and QE$_{up}$, respectively. The second drawback is a potential mismatch between the filter that is used while creating the input image/video and the one that is employed while computing the quality. These mismatches are especially influential if we calculate the quality using a pixel-based method like PSNR. Finally, the third drawback results from ignoring the effect of viewing distance on quality. A well-designed QE must take into account the viewing distance because the sensitivity of the human visual system (HVS) depends heavily on the spatial frequency, which changes with the viewing distance as described in [22].

The third subgroup consists of QEs which are designed to account for the effect of spatial or temporal resolutions. To the best of our knowledge there are not any image QEs in this subgroup. The video QEs approaches described in [23]-[28] belong to the third subgroup. In [23]-[28], the authors extract several features of the video and use regression to find the optimum weights to include the effect of spatial or temporal resolutions. Feghali et al. [23] and Peng et al. [24] developed QEs to measure the joint effect of temporal resolution and quantization. They both use a weighted combination of PSNR and motion information to estimate the quality. In [23], the authors find the threshold corresponding to the largest 25% of the motion vectors, while in [24], the authors employ MPEG-7 motion activity. Both [23] and [24] do not take into account the effect of spatial resolution. Refs. [25] and [26] explore subjective video quality on mobile platforms when spatial resolution, temporal resolution, and quantization parameters are varied. Their results show that the amount of decrease in subjective quality changes according to the content and the quantization parameter when the spatial resolution is decreased. Kim et al. developed a metric in [27] which is a weighted combination of PSNR, motion activity and a logarithmic function of frame size. Sohn et al. [28] extend this work by inserting spatial complexity into the model. Motion activity and spatial complexity features of videos are extracted using MPEG-7 motion activity and MPEG-7 edge histogram functions, respectively. The performance of these estimators are measured in [29] and [30]. In [23]-[28], the scaling factors are obtained using regression. Although the QEs in the third subgroup provide a scientific way to compute the quality, they depend on a specific dataset. This dependence might affect the reliability of these approaches if we apply them to videos with different characteristics. In these studies, the authors also resize either the reference video or the input video while performing the subjective tests.

Having a reliable QE for measuring the quality of visual contents with different spatial resolution is very important. Moreover, it is also critical to have such a QE in performing perceptual optimum video multicasting. Multicasting has been studied thoroughly [31] for wired and wireless networks when the devices have the same resolution and the channels of these devices are heterogeneous. On the other hand, there are a few studies which work on multicasting videos to devices with different spatial resolutions and these studies use scalable video coding. Lee et al. [32] consider the change in the spatial scalability while performing multicasting. In [33], the authors use both spatial and temporal scalabilities. The authors in [34] and [35] employ the combination of SNR and Spatial scalabilities. These authors use scalable video coding in combined scalibility mode which allows to use more than one scalability. They work on the problem of optimum wireless multicasting of scalable video to the users with heterogeneous displays using maximum utility and maxmin fairness constraints, respectively. Although performing optimization for a specific constraint is a straight forward problem, there are two important issues to take into account to do perceptually optimum multicasting using combined scalability. The first one is the dimension increase due to usage of more than one scalability option. The second one is the complexity of the QE used while computing the utility.

1.2 Thesis Overview

In this thesis, we analyze the perceptual quality estimation concepts and apply it in a video delivery scenario for different spatial resolutions. First, we design a quality estimator for images. Secondly, we consider the effect of motion on the quality and extend our design to cover video quality estimation. Finally, we apply the quality estimator concept we develop in a perceptually optimum multicasting scenario. We explain the progress in thesis as follows: In Chapter 2, we describe full reference quality estimation for images with different spatial resolutions. In Chapter 3, we introduce full reference quality estimation for videos with different spatial resolutions. In Chapter 4, we propose an algorithm to reduce the complexity of perceptually optimum multicasting at low bitrates. Chapter 5 concludes the thesis.

Chapter 2

Quality Estimation for Images with Different Spatial Resolutions

In this chapter, we design a QE to estimate the quality of the corrupted displayed image with a lower resolution compared to the reference image. We also propose a subjective test environment where we can evaluate the success of the proposed algorithm in comparison with other QEs. When the reference and test images have different resolutions, the subjective tests demonstrate that in most cases the proposed method works better than other approaches. In addition, the proposed algorithm performs well when the reference image and the test image have the same resolution. Section 2.1 explores the potential weaknesses of previous approaches that are used to estimate the quality in our problem. In Section 2.2, we describe our proposed QE approach, called Multiscale Image Quality Estimator (MIQE) and its colored version. In Section 2.3, we describe the subjective test environment and analyze the results of the subjective tests.

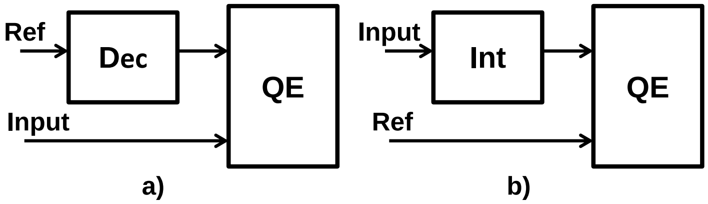


Figure 2.1: Conventional Approaches : a) Input image vs decimated reference image b) Interpolated input image vs reference image

2.1 Limitations of the Previous Approaches

To date, QEs have been designed to estimate quality when the displayed image has the same number of pixels, or spatial resolution, as the reference image. However, viewers use a variety of devices, including TVs and smart phones, to watch movies and perform video conferencing. These devices have different spatial resolutions and are viewed from different distances. To adapt existing QEs to estimate quality when the displayed image has a different spatial resolution, we use QE$_{down}$ and QE$_{up}$ approaches as described in Chapter 1.1. These are illustrated in Figure 2.1a and 2.1b, respectively for the case where the image being considered has lower resolution than the reference image. (The extension to the opposite case is straightforward.) QE$_{down}$ is used in Refs. [4] and [36] , and QE$_{up}$ is used in Refs. [37] and [38] to estimate the quality. Despite their ease of use, QE$_{down}$ and QE$_{up}$ have significant limitations [39]. In this section, we present a motivating example that demonstrates these two approaches fail to accurately estimate subjective quality, then describe their drawbacks and limitations. We begin with some definitions.

2.1.1 Definitions

To mathematically compare two images that are viewed with different resolutions and/or different distances, we describe the images using an angular frequency representation. Specifically, we compute the angular frequency using [40]:

$$f(l) = \frac{\pi * d * n}{180 * h * 2 * 2^{l}} \tag{2.1}$$

In this expression, $f(l)$ denotes the angular frequency in cycles per degree (cyc/deg); d, h, and n represent the distance of the viewer, height of the screen, and the number of pixels in the vertical direction, respectively. In addition, l indicates the level of a subband decomposition. When there is no subband decomposition, we set $l = 1$.

Let the term *input image* describe the low-resolution degraded image, $\mathbf{i}$, that is input to either QE_{down} or QE_{up}. Further, let $\mathbf{r}$ be the *reference image.* We reserve the term *test image* to denote those images that are shown to viewers during a subjective test. The decimated reference image, $\mathbf{d}_r$, is input to the QE-block in Figure 2.1a, while the interpolated input image, $\mathbf{u}_i$, is input to the QE-block in Figure 2.1b.

2.1.2 Motivating Example Using QE_{up} and QE_{down}

Our motivating example, from [39], uses the Visual Information Fidelity (VIF) [41] to compare image compression using JPEG 2000 with and without downsizing the image prior to compression. VIF has been demonstrated to have good performance in a variety of applications. Figure 2.2 shows quality estimated using VIF, as a function of compression level, for the 512×512 "log-seaside" image from the CSIQ database [42], when JPEG 2000 is applied to either a full-sized or half-sized image. The curve labeled *full-size* is the conventional VIF applied to the full-size compressed image. Two methods of computing VIF for the half-size compressed image are also shown. The curve labeled "Half-size (QE_{down})" is VIF computed using QE_{down}, and the curve labeled "Half-size (QE_{up})" is VIF computed using QE_{up}. To create a fair comparison, the compression level for the half-sized images incorporates the impact of downsampling.

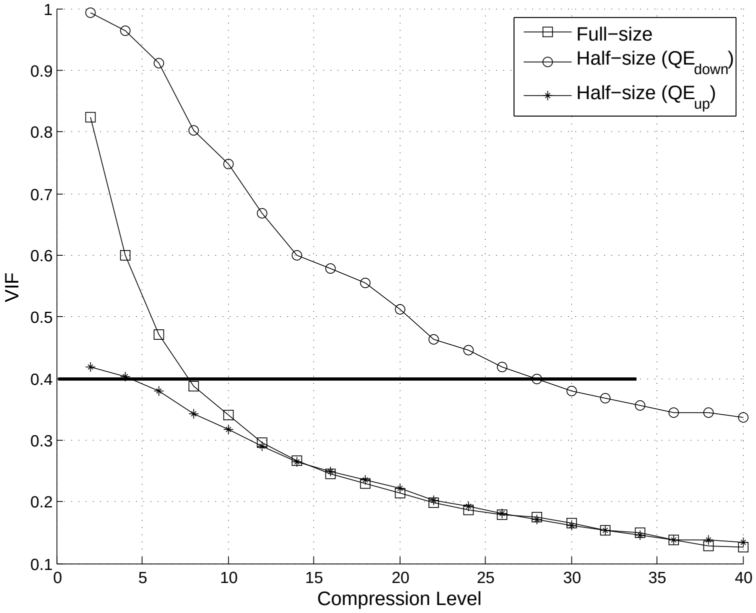


Figure 2.2: Calculated VIF values of the test images for each compression level

To explore whether VIF is accurate or not for this problem, we choose a VIF value of 0.4, corresponding to the dashed line in Figure 2.2. A visual inspection of these three images indicates that the half-size image computed using QE_{up} is substantially better than the full-size image, while the half-size image computed using QE_{down} is worse than both. Therefore, these three images have different visual quality although they have the same estimated quality. More details, including the three images, can be found in [39].

To understand the limitations of QE$_{down}$ and QE$_{up}$ better, we next describe the impact of these methods in the frequency domain.

2.1.3 Frequency Domain Comparison of QE$_{down}$ and QE$_{up}$

To estimate image quality objectively, **r** and **i** should be compared at the same visual angle. Our motivating example indicates that both QE_{up} and QE_{down} are not accurate. To understand their limitations better, we describe the effects of these methods pictorially, using

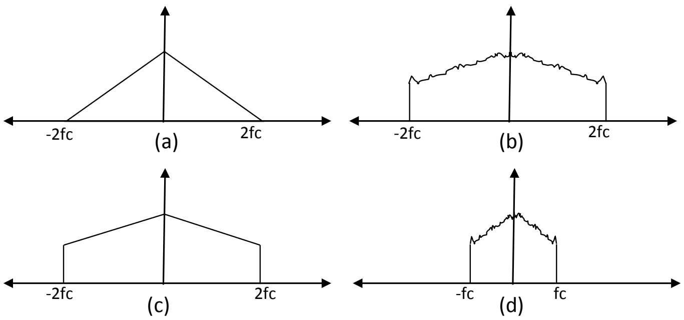


Figure 2.3: 2D Projected Frequency Domain Representations of Image Models (a) Reference image (b) Input Image (c) Decimated Reference Image (d) Interpolated Input image. The downsampling rate is 2

simplified 1D frequency domain representations.

Figure 2.3a shows a sketch of the frequency representation of a *reference image*, $\mathbf{r}$, where the horizontal axis indicates frequency in cycles per degree and the vertical axis shows the magnitude of the frequency content of the signal. While we do not know how the *input image*, $\mathbf{i}$, is generated from the *reference image*, it is reasonable to assume that is obtained by low-pass filtering (LPF), downsampling (DS) , and some additional degradation. Figure 2.3b illustrates the frequency representation of such an *input image*. In this subsection, upsampling and downsampling operations rates are 2.

First, we consider the method of QE_{down}, where we compare a decimated reference image at downsampling rate 2, $\mathbf{d}_r$ with $\mathbf{i}$ as in Figure 2.1a. During QE_{down}, LPF and DS operations are performed consecutively. The resulting decimated reference image in the frequency domain is shown pictorially in Figure 2.3c, where the x-axis is in cycles/degree and y-axis illustrates the magnitude. The QE block of QE_{down} in Figure 2.1a compares the images represented in Figures 2.3c and 2.3b.

Similarly, the method QE_{up} estimates quality of an input image that is smaller than the reference image by comparing an interpolated $\mathbf{i}$ with the reference image $\mathbf{r}$. During QE_{up}, we upsample (US) $\mathbf{i}$ and perform LPF. The frequency domain representation of the interpolated *input image*, $\mathbf{u}_i$, is shown pictorially in Figure 2.3d. The QE block of QE_{up} in Figure 2.1b compares the images represented in Figures 2.3a and 2.3d.

Both QE_{down} and QE_{up} have several drawbacks. In QE_{down}, high-frequency content is lost from the reference image prior to comparison. On the other hand, in QE_{up}, the input image is further degraded before the comparison due to the LPF. In the former case, quality will be estimated as better than the actual quality, because the reference has been degraded. In the latter case, quality will be estimated as worse than the actual quality, because the input image has been further degraded.

A second drawback arises due to potential mismatch between the filters. Since it is not known what filter has been used to create the actual input image $\mathbf{i}$, it is not possible to ensure the same filter is used in QE_{down} to create $\mathbf{d}_r$. Difficulties arise not only due to mismatch between the magnitude responses of the LPFs, but also a spatial shift between the two images might be introduced if the two filters have different phase responses. This could significantly affect the accuracy of the estimated quality.

Finally, one factor that is completely ignored in both QE_{up} or QE_{down} is the viewing distance. Many existing QEs also do not account explicitly for viewing distance. However, the sensitivity of the HVS depends heavily on the frequency. The Contrast Sensitivity Function (CSF) [43] describes how humans perceive spatial frequency, and is effectively the spatial frequency response of the HVS. The CSF has more weight at lower frequencies, and recent subjective tests in [4] and [5] also show that our eyes are more resilient to distortions as the spatial resolution decreases. Upsampling contracts $\mathbf{i}$ in the frequency domain, causing the artifacts in $\mathbf{u}_i$ to become more visible. Therefore, QE_{up} will incorrectly estimate the quality of $\mathbf{i}$.

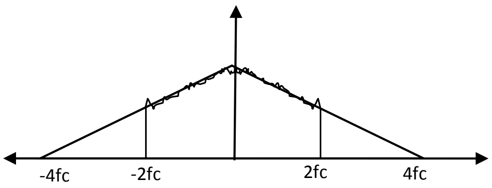


Figure 2.4: Pictorial representation of the reference image and the input image that have the same visual angle in the frequency domain

2.1.4 Design Requirements

To develop an accurate QE for our application, our design must satisfy several requirements. First, our QE must allow the input image **i** and reference image **r** to have different spatial resolutions. This requirement is not satisfied by pixel-based QE methods, like MSE and SAE. To fulfill this requirement, our QE is based on a frequency or wavelet representation. However, this is not sufficient, since as we have seen, even QEs like VIF which use a wavelet decomposition are forced to rely on either QE_{up} or QE_{down} when **i** has a different spatial resolution than **r**.

Second, our design must compare **r** and **i** as if they are being viewed at the same visual angle. To accomplish this, we assume that **r** is placed at a greater distance. Using the angular frequency resolution defined in Eq. (2.1), if we increase the distance by a factor of two, the visual display resolution of **r** is also multiplied by two. The frequency domain representations of the **i** and the relocated **r** are shown in Figure 2.4. In this figure, the downsampling rate is 2. This requirement enables us to adjust the sensitivity of HVS according to the angular resolution. Hence, we model the effect of resolution change to the perception of the distortion accurately.

The final requirement is robustness to small shifts. It is not possible to know the processing that was applied to create the input image **i**. In particular, filtering and downsampling often introduce small shifts into the resulting image. Moreover, using wavelet transform

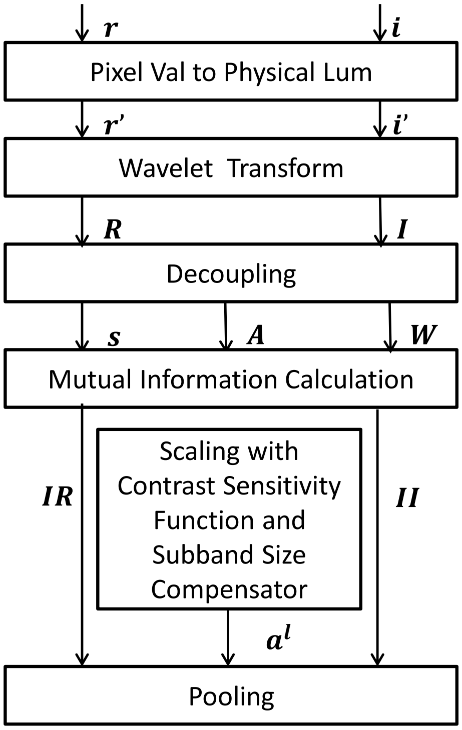

r
i
Pixel Val to Physical Lum
r'
i'
Wavelet Transform
R
I
Decoupling
s
A
W
Mutual Information Calculation
IR
Scaling with Contrast Sensitivity Function and Subband Size Compensator
II
a^l
Pooling

Figure 2.5: Block diagram of the proposed QE

can also cause shift during decomposition. Therefore, for this application of estimating the quality of an image that has a different spatial resolution than the reference image, an accurate QE must be robust to small shifts. To achieve this, our strategy is to compute the *correlation* between image subbands instead of performing point by point comparison between these subbands. In the next section, we describe how we use wavelet decomposition and correlation to develop the proposed algorithm.

2.2 Multiscale Image Quality Estimator

The block diagram of the proposed QE, described in this section, is illustrated in Figure 2.5. As before, **r** represents the reference image and **i** represents the input image. The first block is *Pixel Value to Physical Luminance Conversion.* In this block, pixel values of $\mathbf{r}$ and $\mathbf{i}$ are converted to luminance values and these values are kept in $\mathbf{r}'$ and $\mathbf{i}'$ matrices. The second block is *Wavelet Decomposition.* $\mathbf{r}'$ and $\mathbf{i}'$ are decomposed into their subbands. These bands are collected in $\mathbf{R}$ and $\mathbf{I}$ matrices, respectively. *Decoupling* is the third block. It divides the subbands of $\mathbf{R}$ and $\mathbf{I}$ into smaller non-overlapping blocks. *Decoupling* consists of two steps. In the first step, blocks of $\mathbf{R}$'s subbands are represented using Gaussian Scale Mixture (GSM). In the second step, the blocks of $\mathbf{I}$ are estimated from the blocks of $\mathbf{R}$ using a linear least square estimator. In the fourth block of the diagram, the *Mutual Information* is calculated between the blocks of $\mathbf{R}$ and $\mathbf{I}$. However, each subband has a different importance. Therefore, in the fifth block, we scale the similarity of each subband with the corresponding *sensitivity value and subband size compensator.* Finally, in the last block, *Pooling*, the value of the estimated quality is found. The final quality value is between 0 and 1, and this value increases as the quality increases. In what follows, we explain the details of each block.

Pixel Value to Physical Luminance Conversion

In the first block, we convert pixel values in $\mathbf{r}$ and $\mathbf{i}$ to luminance values $\mathbf{r}'$ and $\mathbf{i}'$. The image information is kept as pixel values in the memory. However, the display changes these values to adapt to HVS. These adapted values are called luminance values. Therefore, it is necessary to alter the pixel values to luminance values. We use the following equation to calculate the corresponding luminance value for each pixel [44]:

$$L(P) = (b + k * P)^{\gamma} \tag{2.2}$$

In this equation, b, k and, γ represent black-level offset, voltage scaling factor and the gamma of the display monitor, respectively. Typical parameter values are $b = 0$, $k = 0.02874$ and $\gamma = 2.2$. We apply this operation to both the reference image, $\mathbf{r}$, and the input image, $\mathbf{i}$, to obtain their luminance counterparts $\mathbf{r}'$ and $\mathbf{i}'$.

Wavelet Transform

In this step, the luminance image matrices $\mathbf{r}'$ and $\mathbf{i}'$ are decomposed into their subbands. There are two reasons for using subband decomposition. First, the space-frequency localization of the HVS is frequently modeled using wavelet-based approaches. Watson [45] modeled it using Cortex Transform. Daly also used a similar transform in Visible Difference Predictor (VDP) [46] model. Second, using a frequency-domain representation is required to enable a scale-independent QE. In this block, we use the 9/7 bi-orthogonal wavelet transform [47]. The main disadvantage of using wavelet transforms is the fact that the ratio between the size of the input image and the size of the reference image can only be a power of 2. Nevertheless, the ratio of popular visual content standard pairs like CIF-QCIF, VGA-QVGA and HDTV-UHDTV satisfy this requirement. The inputs to the wavelet decomposition, $\mathbf{r}'$ and $\mathbf{i}'$, contain the spatial luminance values. Output matrices of the wavelet decomposition block are $\mathbf{R}$ and $\mathbf{I}$. They are the wavelet transforms of $\mathbf{r}'$ and $\mathbf{i}'$, respectively. Matrices $\mathbf{R}$ and $\mathbf{I}$ consist of $3 \times L + 1$ subbands where L represents the total number of levels. Each subband is represented with its level and orientation indices. Specifically, if $\mathbf{I}_{l,o}$ represents a subband of the $\mathbf{I}$, then l and o stand for the level and orientation, respectively. The subbands in the first level of the $\mathbf{I}$, ($\mathbf{I}_{1,2}$**,** $\mathbf{I}_{1,3}$ and $\mathbf{I}_{1,4}$), are $\mathbf{0}$ matrices.

Decoupling

Now, we have the wavelet subbands in $\mathbf{R_{l,o}}$ and $\mathbf{I_{l,o}}$ matrices. In this block, we find the relationship between each subband of $\mathbf{R}$ and $\mathbf{I}$, using two steps. In the first step, $\mathbf{R_{l,o}}$ are represented as Gaussian Scale Mixtures (GSMs). In the second step, blocks of $\mathbf{I_{l,o}}$ are estimated from the blocks of $\mathbf{R_{l,o}}$. To estimate $\mathbf{I_{l,o}}$ blocks, we assume that there is a linear

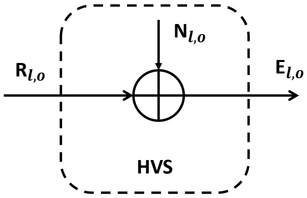


Figure 2.6: Mathematical model of the reference image observation

relationship between the corresponding blocks of $\mathbf{R_{l,o}}$ and $\mathbf{I_{l,o}}$. We represent this linear relationship with attenuation and additive noise. We also assume that the additive noise is White Gaussian Noise and independent of the blocks of $\mathbf{R_{l,o}}$. Hence, we can use a Linear Least Square Estimator (LLSE) to find these estimation parameters. We elaborate the details of the decoupling process in the following paragraphs.

The first part of the decoupling process is based on a statistical evaluation of images. In this part, we assume that the reference image is accurately characterized by a model based on natural scene statistics (NSS) [48]. The hypothesis of NSS is that images of natural scenes occupy only a small subset of all possible images. Wainwright et al. [49] have used this information to represent images in terms of GSMs. To do that, they represent each subband of the image as a summation of Gaussian distributions [49]:

$$\vec{R}_{l,o,j} = S_{l,o,j}.\vec{U}_{l,o,j}, j = 1, \ldots, K \tag{2.3}$$

The index of the block is shown by j, and it can take values from 1 to the number of the blocks in the subband. In this expression, $\vec{R}_{l,o,j}$ represents the j^{th} block of subband (l,o). $\vec{U}_{l,o,j}$ is a Gaussian vector with mean zero and covariance $\mathbf{C_u}$. The covariance $\mathbf{C_u}$ can be computed by averaging the autocorrelation of $\vec{U}_{l,o,j}$ in the same subband. $S_{l,o,j}$ is a random number from a positive scalar random field. Given $S_{l,o,j}$, $\vec{R}_{l,o,j}$ is normally distributed with $N(0, S^2_{l,o,j}\mathbf{C_u})$.

Hence, after finding $S_{l,o,j}$, the blocks can be characterized as Gaussian distributions.

The aim of the second part of the decoupling process is to represent the input blocks, $\vec{I}_{l,o,j}$, in terms of the reference blocks, $\vec{R}_{l,o,j}$. Here, the noise is assumed to be orthogonal to $\vec{R}_{l,o,j}$, so we use the LLSE to find the attenuation ($A_{l,o,j}$) and additive noise ($\vec{W}_{l,o,j}$) parameters using the following equations.

$$\vec{I}_{l,o,j} = A_{l,o,j}\vec{R}_{l,o,j} + \vec{W}_{l,o,j} \tag{2.4}$$

$$A^*_{l,o,j} = \min_{A_{l,o,j}} ||\vec{I}_{l,o,j} - A_{l,o,j}\vec{R}_{l,o,j}||_2 \tag{2.5}$$

This process is performed for each $\vec{R}_{l,o,j}$-$\vec{I}_{l,o,j}$ block pair. In Eq. (2.5), $A^*_{l,o,j}$ denotes the optimum attenuation that minimizes the noise. We can find the optimum attenuation by setting the derivative of Eq. (2.5) to $\mathbf{0}$. Next, we obtain the variance of the additive noise by subtracting the attenuated reference signal from noise. The corresponding attenuation and noise variance terms are expressed as follows:

$$A_{l,o,j} = cov(\vec{I}_{l,o,j}, \vec{R}_{l,o,j})var(\vec{R}_{l,o,j})^{-1} \tag{2.6}$$

$$var(\vec{W}_{l,o,j}) = var(\vec{I}_{l,o,j}) - A_{l,o,j}cov(\vec{I}_{l,o,j}, \vec{R}_{l,o,j}) \tag{2.7}$$

In Eq. (2.6), $\vec{R}_{l,o,j}$ is a 16x1 vector. Hence, $var(\vec{R}_{l,o,j})$ is a scalar and we can find the inverse as long as $var(\vec{R}_{l,o,j})$ is not zero. To prevent having a zero in the denominator, we add a very small regularization constant to the $var(\vec{R}_{l,o,j})^{-1}$. These attenuation and noise variance values are obtained for each block, and are used in the next step to find the distortion in each block, employing mutual information as the distortion metric.

Mutual Information

Now, we have subband blocks of **I** and **R** in terms of GSMs. We use a similar model mentioned in the mutual information calculation of VIF [41]. Mutual information is a function of two random variables and is an indication of how much information we get from one random variable about the other one. As such, a lower mutual information is an indication of lack of similarity. In the extreme case of zero mutual information, the two random variables are independent. Therefore, mutual information of each block gives us an indication of local similarity.

R and **I** are the reference and input signals. We model **E** and **F** as being created by passing **R** and **I** through the HVS, as in [41], where the HVS is modeled as Additive White Gaussian Noise (AWGN) with unit variance. $A_{l,o,j}$ and $\vec{W}_{l,o,j}$ represent the attenuation, noise for each subband block, respectively. $s_{l,o,j}$ is a realization of $S_{l,o,j}$ and it is computed as in [41], Eq. (15). In this step, two items will be performed as follows.

First, the mutual information [50] between non-overlapping 4x4 blocks of $\mathbf{E_{l,o}}$ and $\mathbf{R_{l,o}}$ ($I(\mathbf{E_{l,o}};\mathbf{R_{l,o}})$) is found for all (l,o) pairs. Since we have used a correlation based block similarity metric during the mutual information computation, we have relieved the effect of shift variance which can occur as a result of wavelet decomposition or decimation filter mis-match. $\mathbf{E_{l,o}}$ is obtained by adding $\mathbf{R_{l,o}}$ with **N** which is an AWGN, as illustrated in Figure 2.6. Before computing the mutual information, the block indices are represented as a vector. Moreover, $\vec{R}_{l,o,j}$ has a Gaussian distribution if $\mathbf{s_{l,o,j}}$ is known. Using these facts, we can find the mutual information for a (l,o) pair as

$$I(\mathbf{E_{l,o}};\mathbf{R_{l,o}}|s^K) = \frac{1}{2}\sum_{j=1}^{K} log_2\left(\frac{s_{l,o,j}^2\mathbf{C_u} + \sigma_n^2\mathbf{I}}{\sigma_n^2\mathbf{I}}\right) \tag{2.8}$$

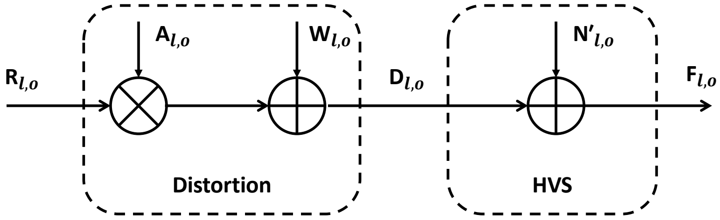


Figure 2.7: Mathematical model of the input image observation

Second, the mutual information between the blocks of $\mathbf{F_{l,o}}$ and $\mathbf{R_{l,o}}$ is found. The relationship between $\vec{F}_{l,o,j}$ and $\vec{R}_{l,o,j}$ is shown in Figure 2.7. Using this relationship we can write $\vec{F}_{l,o,j}$ in terms of $\vec{R}_{l,o,j}, A_{l,o,j}, \vec{W}_{l,o,j}$ and $\vec{N}'_{l,o,j}$ as follows:

$$\vec{F}_{l,o,j} = A_{l,o,j} * \vec{R}_{l,o,j} + \vec{W}_{l,o,j} + \vec{N}'_{l,o,j} \tag{2.9}$$

We use Eqs. (2.8) and (2.9) to find the mutual information between $\mathbf{F_{l,o}}$ and $\mathbf{R_{l,o}}$ as follows:

$$I(\mathbf{F_{l,o}}; \mathbf{R_{l,o}}|s^K) = \frac{1}{2}\sum_{j=1}^{K} log_2\left(\frac{s^2_{l,o,j}A^2_{l,o,j}\mathbf{C_u} + \sigma^2_{q_{l,o,j}}\mathbf{I}}{\sigma^2_{q_{l,o,j}}\mathbf{I}}\right) \tag{2.10}$$

where, $\sigma^2_{q_{l,o,j}} = \sigma^2_{w_{l,o,j}} + \sigma^2_n$. We use Eqs. (2.8) and (2.10) to find the mutual information for each subband. However, we also need to scale these values ($I(\mathbf{E_{l,o}}; \mathbf{R_{l,o}})$, $I(\mathbf{F_{l,o}}; \mathbf{R_{l,o}})$) according to the size and the importance of the subband.

Scaling with Contrast Sensitivity Function and Subband Size Compensator

To estimate the quality accurately, we next scale the value of each (l, o) pair. To calculate the magnitude of these scaling coefficients, we take into account two factors: the visual importance of each subband determined by the characteristics of HVS, and the size difference of each subband.

First, HVS has a different weight for each subband [43]. The corresponding weight changes according to the resolution and height of the screen, the distance between the viewer and the screen, and the level of the subband. To find the effect of HVS, we first calculate the angular frequency by Eq. (2.1). Next, the CSF is calculated using the following formula from [40]:

$$CSF(f) = (0.69 + 0.31 * f).e^{-0.28*f} \tag{2.11}$$

This CSF function is used to compute the quality of monochromatic images. We can take into account the effect of color if we use the chromatic and achromatic CSF functions as described in [51]. In calculating the QE, if we use both chromatic and achromatic CSF functions instead of the one described in Eq. (2.11) the corresponding QE is called MIQE-Color (MIQEC).

The second factor is the size of the subband. As the subband level increases by one, the size of the band decreases by four. Rouse et al. [52] show that incorporating this improves the accuracy of quality estimation. As a result, each subband CSF value is also scaled with 2^{2l} to compensate the size of each level l. Hence, the scaling coefficient for each subband is calculated as follows:

$$g_l = CSF(f(l)) * 2^{2l} \tag{2.12}$$

Pooling

In this step, we calculate the quality using the scaling coefficients and mutual information for each subband. Specifically, we compute the estimated quality by

$$MIQE = \frac{\sum_{l,o \in subbands} g_l * I_{I,l,o}}{\sum_{l,o \in subbands} g_l * I_{R,l,o}} \tag{2.13}$$

In this expression, g_l is the scaling coefficient for level l given in Eq. (2.12). $I_{I,l,o}$ and $I_{R,l,o}$ represent the scaled mutual information of the input and the reference signals for the pair (l, o).

To check the validity of the proposed algorithm, it is necessary to compare the estimated quality values with the preference of viewers. In the following section, we describe the details of our subjective tests to obtain viewer preferences.

2.3 Subjective Tests

Subjective tests are performed to obtain the viewers' preferences for images that have different spatial resolutions. The test results are used both to ensure that the proposed algorithm is accurate and to compare our algorithm with existing approaches. Our subjective tests are performed by employing a paired comparison test using two distorted test images, one with full resolution and the other with half resolution. In the following subsections, we describe how we create the test set, perform subjective tests, and analyze results.

2.3.1 Test Set Creation

Original images are selected from the National Park Service Digital Image Archive and WikiMedia Commons. There are in total 24 images in 5 categories. These categories are human, landscape, plant, animal, and urban. These images represent a wide variety of scenes. They include homogeneous regions, edges, and details. The resolution of the reference images is 800×800. The images that are downloaded from the databases have greater resolution than 800×800; hence, they are decimated and cropped. We decimate the images such that minimum edge resolution of the decimated image is greater than or equal to 800. Next, we determine the border coordinates to crop the image. We choose these coordinates to keep the maximum information in the reference image.

The test images shown to viewers consist of two groups: high resolution and low resolution images. High resolution test images are obtained by distorting the reference images. Low resolution test images are created by decimating the reference images and then corrupting the decimated images.

To obtain a subjective estimate of image quality, it is conventional to ask viewers for their opinion score and then compute the mean opinion scores (MOSs). This method is effective when the reference image and the test image have equal resolutions; however, finding the qualities of images that have different spatial resolutions is more challenging [6]. Therefore, instead of obtaining the MOS score for each image, we use paired comparison to find the relative preference of test images with different resolutions.

To obtain the test images, we corrupt a full-size reference image $\mathbf{r}$ and a decimated reference image $\mathbf{d}_r$ using one of the following four distortions: compression with JPEG, compression with JPEG 2000, blurring, or noise. For the two compression algorithms, we use three distortion levels, representing good quality, medium quality and bad quality images. However, both blur and noise have only two distortions levels: good and bad. A $2D$ Gaussian low pass filter is used for blurring, where the distortion level of the blur is controlled by the standard deviation of the filter. In addition, we use AWGN as noise distortion. To explore the impact of filter type, we use two low-pass filter types to decimate the $\mathbf{r}$. The first is the Non-Normative LPF, which is a Sine-windowed Sinc-function. We implement the filter employing Eq. (4) from Segall and Sullivan [53]. The parameters are $D = 2.5$ and $N = 3$ and phase offset$= 0$, and the implementation is performed using floating point. The second LPF is the Raised-Cosine function which is formulated as follows.

$$f(x) = \frac{1}{2}(1 + cos(\frac{\pi * (w - w_c * (1 - \alpha))}{2 * \alpha * w_c})) \tag{2.14}$$

In this expression, w_c is the cut-off frequency and α is the roll-off factor. Non-Normative filter only causes blurring. It especially decreases the magnitude of the signal between frequencies 0.4π and 0.5π. On the other hand, Raised Cosine causes both aliasing and blurring. When it is necessary to distinguish the two decimated reference images, we use $\mathbf{d}_{r1}$ and $\mathbf{d}_{r2}$, to represent $\mathbf{d}_r$s created using the Non-Normative and Raised-Cosine LPFs, respectively. All distortion types are used to distort decimated reference images $\mathbf{d}_{r1}$, while only JPEG 2000 is used to distort decimated reference images $\mathbf{d}_{r2}$. We use JP2K-RC as an abbreviation for the latter case. To minimize the number of comparisons in our subjective test, we use all images in the reference image set to obtain $\mathbf{d}_{r1}$ and half of the images in the set to obtain $\mathbf{d}_{r2}$.

The next challenge is to determine appropriate levels of distortion for each image pair in the paired comparison experiment. Ideally, we should choose the distortion level of the low resolution test image, relative to that of the high resolution test image, such that we obtain the most information about how the QEs perform. If we choose the image pairs such that all QEs rank the images in the same order, our subjective test will not be able to distinguish the relative performance among the QEs. If instead, we pick an image pair such that one QE classifies the images correctly and the remaining QEs classify them incorrectly, we immediately determine the best QE for that image pair. However, this is difficult because we do not know the viewers' preferences a-priori.

In consequence, we use a method similar to that described in [54] to select the images. According to this approach, for a given high resolution test image, we select a corresponding low resolution test image which divides a set of QEs into two groups, with approximately half of the QEs classifying an image pair one way and approximately half classifying the pair the other way. We perform QE$_{down}$ approach while estimating the quality of the low resolution test image using the Non-Normative filter [53].

We use this strategy to create a pair of images for each reference image, each distortion, and each distortion level (good, medium, or bad), where one image in the pair has high resolution and the other low. In our subjective test, in addition to simply having viewers rate this pair, we create additional pairs using these images. For JPEG and JPEG 2000, we use three additional pairs per reference image by pairing the bad-quality high-resolution image with the medium-quality low-resolution image, and the medium-quality high-resolution image with both good and bad low-resolution images. For blur and noise, we create one additional pair per reference image by pairing the bad-quality high resolution image with the good-quality low resolution image. These additional pairs provide additional information about viewers' preferences while balancing the expense of the subjective test.

2.3.2 Subjective Test Implementation

Until now, we have described how we create the test images and form the possible pairs. Next, we describe the subjective tests themselves. The tests contain a total of 504 pairs of images, with 60 people viewing 252 pairs each. Hence, each pair is seen by 30 viewers, and each test session takes approximately an hour. The viewers selected for the test are graduate students and undergraduate students. They have clear vision and they are non-experts.

During tests, the viewers saw images that have different spatial resolutions side by side. They were asked to choose which image they prefer more. They had 10 seconds to see each image pair, but they could choose a preference before or after 10 seconds. The distance between the viewers and the screen was 4 times the height of the low resolution image. Tests were performed using 17" Dell N7110 Inspiron with a display resolution of 1600 × 900.

2.3.3 Analysis of Subjective Test Results

In this section, we compare the performance of our proposed QE to nine others using the results of the subjective test. We begin by examining whether the QEs can rank the images in each pair according to the viewers' preferences using QE_{down} setup. Then, we explore the sensitivity to the choice of decimation filter applied in the QE_{down}, with emphasis on the case when it differs from the filter used to create low resolution test images. We also study how using the QE_{up} approach will affect the ranking. Next, we apply a well-known statistical model to transform viewer preferences into an estimate of subjective quality among all the images with the same reference image and distortion type. Lastly, we explore the performance of the proposed algorithm when the images have the same resolutions. We compare our proposed MIQE and MIQEC to seven FR and two NR state-of-the-art QE methods. Compared state-of-the-art FR QE methods are Structural Similarity Index (SSIM) [18], VIF [41], Visual Signal-to-Noise Ratio (VSNR) [44], PSNR, Multiscale-SSIM (MSSIM) [19], Information Content Weighted SSIM (IWSSIM) [55] and Complex Wavelet SSIM (CWSSIM) [56]. The two NR QE methods are BIQI [16] and BRISQUE [17].

In our first comparison, we begin by identifying the image that was preferred by most viewers, for each pair shown to viewers. We also use each QE to rank the two images in each pair. Dividing the total number of correct rankings by the total number of pairs gives us the fraction of correct rankings. Table 2.1 shows the percentage of correct rankings for each QE according to different distortions and different decimation filters.

In "Different Distortions" heading of Table 2.1, we evaluate the effect of different distortions to the correct ranking. The Non-Normative filter is used while performing QE$_{down}$, and test images are obtained by distorting the images in $\mathbf{d}_{r1}$. According to the table, MIQEC algorithm performs better than other QEs for all distortions except JP2K. If the distortion is JP2K, MIQE performs better than other QEs. Both MIQE and MIQEC have higher

Table 2.1: Fraction of Correct Rankings

	Different Distortion				Different Filter	
	Blur	JP2K	JPEG	Noise	JP2K	JP2K-RC
SSIM	0.597	0.601	0.569	0.354	0.569	0.556
VIF	0.556	0.531	0.597	0.396	0.500	0.417
VSNR	0.722	0.677	0.667	0.660	0.653	0.653
MIQE	0.819	**0.774**	0.729	0.910	**0.736**	**0.792**
PSNR	0.708	0.747	0.688	0.910	0.708	0.708
MSSIM	0.597	0.524	0.569	0.340	0.500	0.403
IWSSIM	0.542	0.531	0.569	0.340	0.500	0.403
MIQEC	**0.917**	0.760	**0.736**	**0.924**	0.736	0.708
CWSSIM	0.736	0.760	0.639	0.507	0.514	0.583
BIQI	0.667	0.573	0.59	0.563	0.653	0.681
BRISQUE	0.625	0.510	0.576	0.507	0.542	0.514

ranking fractions for blur and noise compared to other distortions. The results also show that the proposed QE is robust to small shifts. The proposed QE uses a wavelet LPF during subband decomposition instead of employing the Non-Normative filter to estimate quality. Nevertheless, MIQEs's correct ranking fraction is higher than other QEs.

In "Different Filters" heading of Table 2.1, the effect of different filters to the correct ranking is analyzed. The fifth and the sixth columns of Table 2.1 illustrate the fraction of correct rankings when the decimation filter is Non-Normative and Raised Cosine LPF, respectively. In each case, we use the same LPF during the creation of test images and the decimation operation in QE$_{down}$. Here, the comparison is performed for the reference images which are common in $\mathbf{d}_{r1}$ and $\mathbf{d}_{r2}$. Note that, the number of images in $\mathbf{d}_{r1}$ and $\mathbf{d}_{r2}$ are different. Therefore, the QE values for JP2K in "Different Distortions" and "Different Filters" are slightly different. The ranking fractions show that MIQE performs slightly better when the Raised Cosine LPF is employed instead of the Non-Normative to create test images. Furthermore, the ranking fractions of other QEs either decrease or remain the same when the Raised Cosine LPF is employed during QE$_{down}$.

We also study the effect of a mis-match between the decimation filters that are utilized in the creation of the test images and QE_{down}. We employ the 9/7 biorthogonal wavelet, raised cosine and non-normative filters to compute the fraction of correct ranking scores in this

Table 2.2: Fraction of Correct Rankings When the Low Pass Filter is a Wavelet Filter

	Different Distortion				Different Filter	
	Blur	JP2K	JPEG	Noise	JP2K	JP2K-RC
SSIM	0.819	**0.788**	0.722	0.590	**0.750**	**0.792**
VIF	0.639	0.545	0.618	0.576	0.528	0.569
VSNR	0.597	0.677	0.729	0.882	0.625	0.750
MIQE	0.819	0.774	0.729	0.910	0.736	0.792
PSNR	0.819	0.719	0.694	**0.965**	0.681	0.764
MSSIM	0.875	0.615	0.715	0.396	0.597	0.528
IWSSIM	0.875	0.642	**0.771**	0.521	0.625	0.528
MIQEC	**0.917**	0.760	0.736	0.924	0.736	0.728
CWSSIM	0.778	0.490	0.556	0.410	0.514	0.486
BIQI	0.667	0.573	0.59	0.563	0.653	0.681
BRISQUE	0.625	0.510	0.576	0.507	0.542	0.514

scenario. If the test images are created using the image set d_{r1}, we apply the wavelet and the raised cosine filters in the decimation block of the QE_{down}. Otherwise, we employ the Non-Normative and Wavelet filters to compute the quality. The filter mis-match may cause the estimated quality to be under-estimated if the QE is not shift invariant. On the other hand, the Dec block causes QE_{down} to over-estimate quality because it causes detail loss in the reference image. The two effects may, coincidentally, cancel each other, producing an accurate quality estimation. No one filter in QE_{down} is likely to perform best for all possible content, distortions, and image-creation filter. When we search for the best decimation filter in QE_{down} over a variety of scenarios, in a few cases, QE_{down} outperforms MIQE and MIQEC. However, MIQE and MIQEC do not rely on a specific filter, and there is no one quality estimation approach that consistently outperforms the others, i.e., the approach providing the highest correct ranking scores varies depending on the applied filter, content, and distortion. On average, MIQEC has the highest ranking scores. Tables 2.2 and 2.3 show the percentage of correct rankings for each QE using Wavelet and Raised Cosine decimation filters in QE_{down}, respectively.

As described in Section 2.1, we can also use QE_{up} to estimate the quality using existing approaches. We use bilinear and 9/7 biorthogonal wavelet filters while interpolating the test images in QE_{up}. We explore the effect of QE_{up} using different distortions and decimation

Table 2.3: Fraction of Correct Rankings When the Low Pass Filter is a Raised Cosine Filter

	Different Distortion				Different Filter	
	Blur	JP2K	JPEG	Noise	JP2K	JP2K-RC
SSIM	0.819	0.684	0.681	0.465	0.667	0.694
VIF	0.514	0.538	0.604	0.465	0.514	0.597
VSNR	0.514	0.747	**0.743**	0.951	0.694	**0.806**
MIQE	0.819	**0.774**	0.729	0.910	**0.736**	0.792
PSNR	0.750	0.760	0.708	**0.951**	0.722	0.792
MSSIM	0.514	0.531	0.611	0.354	0.500	0.500
IWSSIM	0.625	0.545	0.667	0.382	0.514	0.514
MIQEC	**0.917**	0.760	0.736	0.924	0.736	0.728
CWSSIM	0.778	0.503	0.590	0.424	0.514	0.486
BIQI	0.667	0.573	0.59	0.563	0.653	0.681
BRISQUE	0.625	0.510	0.576	0.507	0.542	0.514

Table 2.4: Fraction of Correct Rankings for QE_{up} When Upsampling Filter is Wavelet Filter

	Different Distortion				Different Filter	
	Blur	JP2K	JPEG	Noise	JP2K	JP2K-RC
SSIM	0.486	0.670	0.597	0.924	0.681	0.764
VIF	0.514	0.691	0.597	0.660	0.694	0.750
VSNR	0.694	0.649	0.597	0.660	0.667	0.750
MIQE	0.819	**0.774**	0.729	0.910	**0.736**	**0.792**
PSNR	0.597	0.670	0.604	0.910	0.681	0.750
MSSIM	0.514	0.531	0.611	0.354	0.694	0.750
IWSSIM	0.639	0.684	0.625	0.757	0.681	0.736
MIQEC	**0.917**	0.760	**0.736**	**0.924**	**0.736**	0.728
CWSSIM	0.730	0.670	0.640	0.765	0.684	0.742
BIQI	0.667	0.573	0.59	0.563	0.653	0.681
BRISQUE	0.625	0.510	0.576	0.507	0.542	0.514

filters. For low resolution test images, a QE_{up} approach results in smaller QE values compared to QE_{down}. Ranking classification results in Tables 2.4 and 2.5 show that, MIQE and MIQEC have the highest ranking scores. Moreover, the fraction of correct ranking scores obtained using the wavelet filter are higher than that of the bilinear filter.

The ranking classification only considers whether one image is preferred to the other, either objectively or subjectively, but does not consider the magnitude of the preference. Therefore, in our second comparison we compute the distance similarity between the subjective results and the QE scores. Before describing how we compute distance similarity, we first describe how we obtain distances both subjectively and objectively.

Assigning a continuous quality value to each image is a well-defined problem in statistics

Table 2.5: Fraction of Correct Rankings for QE_{up} When Upsampling Filter is Bilinear Filter

	Different Distortion				Different Filter	
	Blur	JP2K	JPEG	Noise	JP2K	JP2K-RC
SSIM	0.486	0.642	0.597	0.674	0.653	0.764
VIF	0.486	0.642	0.597	0.660	0.653	0.764
VSNR	0.486	0.642	0.597	0.660	0.653	0.764
MIQE	0.819	**0.774**	0.729	0.910	**0.736**	**0.792**
PSNR	0.486	0.642	0.597	0.729	0.653	0.764
MSSIM	0.486	0.642	0.597	0.799	0.653	0.764
IWSSIM	0.486	0.642	0.597	0.729	0.653	0.764
MIQEC	**0.917**	0.760	**0.736**	**0.924**	0.736	0.728
CWSSIM	0.556	0.670	0.729	0.729	0.653	0.764
BIQI	0.667	0.573	0.59	0.563	0.653	0.681
BRISQUE	0.625	0.510	0.576	0.507	0.542	0.514

literature [57],[58]. The relationship between the actual qualities of images I_i and I_j, and the viewers' preference is expressed as follows:

$$P(Y = \pi_{ij}) = Q_i - Q_j + Z \tag{2.15}$$

where Q_i and Q_j denote the qualities of I_i and I_j, respectively, Y is a random variable which represents the tendency to choose I_i over I_j, and π_{ij} denotes the ratio of preferring I_i to all comparisons between I_i and I_j, and Z is a random variable that models the deviation from the actual quality difference. The distribution of Z changes according to the model used. Thurstone-Mosteller (TM) [57] assumes that Z has a Gaussian distribution, whereas Bradley-Terry (BT) [58] models Z with a logistic distribution. Handley [59] shows that the BT model provides two important advantages. First, BT offers a mathematically developed model where quantities like confidence can be calculated analytically. Second, unlike *TM*, we can use BT when we do not have all the paired comparison results [59]. Because of these reasons, we choose BT to convert paired comparisons to continuous values.

We use Eq. (2.15) to find the relationship between the preference probability and quality as follows:

$$\pi_{mn} = \int_{Q_n - Q_m}^{\infty} \frac{e^{-\frac{(x-\mu)}{s}}}{s(1+e^{-\frac{(x-\mu)}{s}})^2} dx = \frac{e^{\frac{\mu}{s}}}{e^{\frac{\mu}{s}} + (\frac{\pi_n}{\pi_m})^{\frac{1}{s}}} \tag{2.16}$$

In Eq. (2.16), $Q_m - Q_n = log(\frac{\pi_m}{\pi_n})$, where π_k's are positive real numbers and $\sum_{k=1}^{N} \pi_k = 1$. μ denotes the mean of the logistic distribution and s is a parameter proportional to the standard deviation of the distribution. If $\mu = 0$ and $s = 1$, we can write $\pi_{mn} = \frac{\pi_m}{\pi_m + \pi_n}$. We find π_m values using maximum likelihood. BT assumes that all comparisons are independent. Hence, we can calculate the likelihood function for all comparisons by.

$$L(\pi_{mn}) = \prod_{m<n} \binom{n_{mn}}{\alpha_{mn}} \pi_{mn}^{\alpha_{mn}} (1 - \pi_{mn})^{n_{mn} - \alpha_{mn}} \tag{2.17}$$

In this expression, n_{mn} denotes the total number of viewers and α_{mn} represents the number of viewers that prefer I_m. To maximize the likelihood, we take the derivative of Eq. (2.17) with respect to π_m and equate it to zero:

$$\sum_{m \neq n} \frac{n_{mn}}{\pi_m + \pi_n} - \sum_{m<n} \frac{\alpha_{mn}}{\pi_m} = 0 \tag{2.18}$$

The BT model provides relative scores only for a specific image and distortion in our paired comparisons. Therefore, we cannot analyze these scores across images or across distortions. Further, only their differences are well-defined, not their actual values. Therefore, we compute the differences of the BT scores, (ΔBT), for all images with the same reference image and same distortion, regardless of spatial resolution.

Next, to determine the distances between these images objectively, we recognize that the QEs have different ranges and often have a nonlinear relationship with subjective quality. Therefore, it is important to normalize the QE scores. In this text, we choose to normalize the QE scores using the observed probability distribution of QE scores across the set of 504

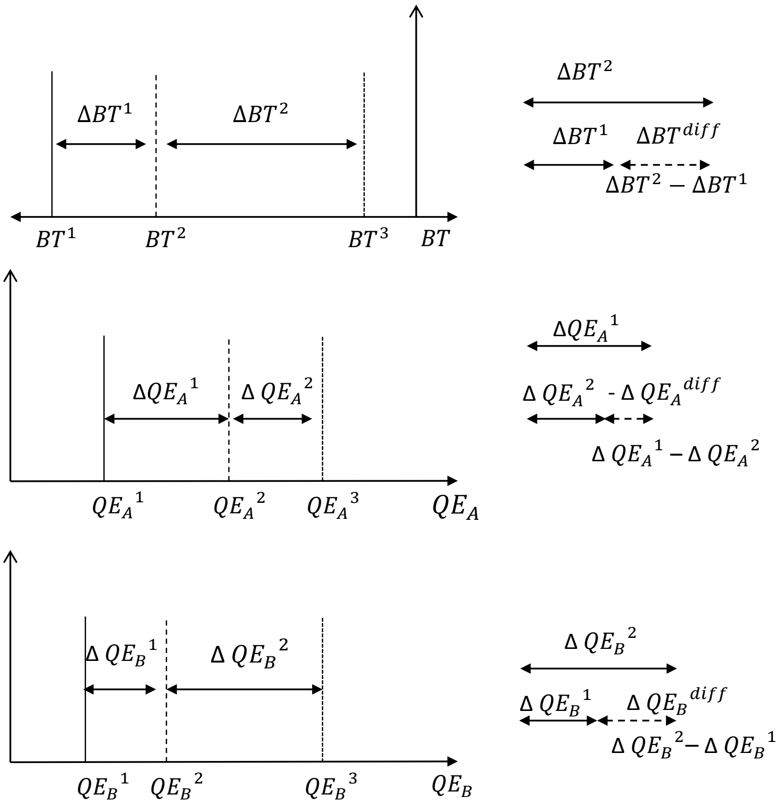


Figure 2.8: Illustration of distance dissimilarity

images we considered. Specifically, we use the observed cumulative distribution function of the QE scores to map the QE scores into values between 0 and 1. During this mapping, we decrease the Kullback-Liebler Divergence [50] among different QEs. We then define an objective distance between two images by the difference, ΔQE, between their normalized QE scores.

Now, we have sets of N images for each reference image and distortion combination, and we have meaningful BT scores defined within each set. We compute the distances between each pair of images, both objectively and subjectively to create a collection of $T = \binom{N}{2}$ pairs. Thus, there are $\binom{T}{2}$ possible combinations of two pairs, which may or may not contain one image in common. To assess the accuracy of the QEs, we compute the similarity of the distances of these combinations.

While there are many ways to compute the similarity among distances, in this text we focus on the following strategy, which is illustrated by the example in Figure 2.8. We have three test images, TI_1, TI_2 and TI_3, each created from the same reference image using the same distortion. The BT scores of the images are BT^1, BT^2 and BT^3, and we assume that as the index of the image increases the quality improves. Therefore, $\Delta BT^1 < \Delta BT^2 < \Delta BT^3$, and $\Delta BT^1 - \Delta BT^2 < 0$ and $\Delta BT^2 - \Delta BT^3 < 0$.

Next, consider two QEs, QE_A and QE_B, for estimating quality, where the QE_A and QE_B values for image TI_i are QE_A^i, and QE_B^i, respectively. As illustrated in Figure 2.8, the distribution of QE_B scores are more similar to the distribution of BT scores than the QE_A scores are. This is because $|\Delta QE_B^1| < |\Delta QE_B^2|$ and $|\Delta BT^1| < |\Delta BT^2|$, while $|\Delta QE_A^1| > |\Delta QE_A^2|$. A correct ordering of distances only happens when all of the following conditions hold: $sign(\Delta BT_1) = sign(\Delta QE_1)$, $sign(\Delta BT_2) = sign(\Delta QE_2)$, and $sign(\Delta BT_1 - \Delta BT_2) = sign(\Delta QE_1 - \Delta QE_2)$.

We employ this approach to compute the distance similarities for each distortion type. We

Table 2.6: Fraction of Correct Distance Similarity For Different Distortions

	Different Distortion				Different Filter	
	Blur	JP2K	JPEG	Noise	JP2K	JP2K-RC
SSIM	0.117	0.369	0.350	0.206	0.344	0.320
VIF	0.264	0.334	0.423	0.161	0.294	0.354
VSNR	0.161	0.366	0.463	0.194	0.392	0.417
MIQE	**0.539**	**0.465**	**0.539**	**0.533**	**0.471**	**0.534**
PSNR	0.167	0.426	0.460	0.161	0.406	0.421
MSSIM	0.147	0.334	0.389	0.192	0.287	0.330
IWSSIM	0.272	0.358	0.400	0.208	0.314	0.343
MIQEC	0.519	0.457	0.469	0.506	0.474	0.342
CWSSIM	0.406	0.328	0.411	0.228	0.324	0.192
BIQI	0.025	0.048	0.019	0.008	0.029	0.103
BRISQUE	0.011	0.052	0.022	0.031	0.058	0.042

also calculate the distance similarities for JP2K when we use different decimation filters during QE computation. We obtain the fraction of correct distance similarities by dividing the total number of correct distance similarities by the total number of comparisons. Table 2.6 presents these fractions for each distortion type and for different decimation filters when the distortion is JP2K. According to the table, MIQE has the highest distance similarity fraction. When the decimation filters are different, the distance similarity values of JP2K-RC is relatively higher than JP2K for all QEs except for SSIM. Hence, it can be inferred that the filter type affects the correct distance similarity fraction.

In addition to these metrics, we explore the relationship between QE score differences and BT score differences graphically. We also compute Spearman Rank Correlation Coefficients (SRCCs) for each QE. Figures 2.9-2.12 illustrate the relationship between four QE scores and BT scores for all distortions. We have picked these QEs because they have the highest SRCC values among all the QEs. The captions of the figures show the SRCC values of the utilized QEs.

Finally, MIQE and MIQEC have been designed to compare images with different spatial resolutions. However, it also has a reasonable performance if the images have the same resolution. We employ LIVE [60], TID [61] and CSIQ [42] databases to evaluate the performance of MIQE and MIQEC when the test and reference images have the same resolution.

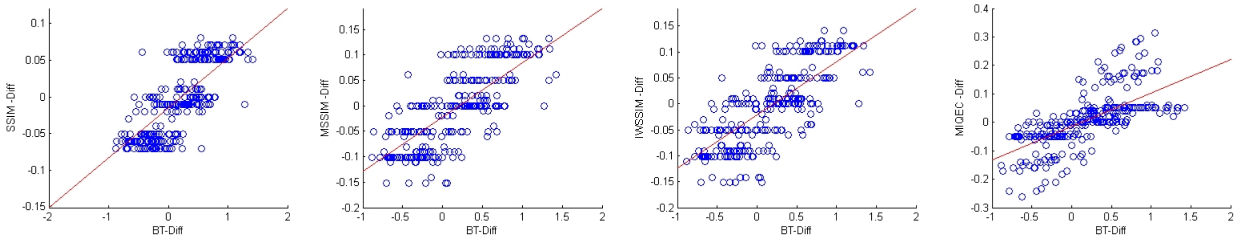


Figure 2.9: Illustration of relationship between QE scores' difference and BT scores' difference when the distortion is Blur. The QEs and their corresponding Spearman Rank Correlation Coefficients are : a) SSIM (0.721), b) MSSIM (0.746), c) IWSSIM (0.724) and d) MIQEC (0.782)

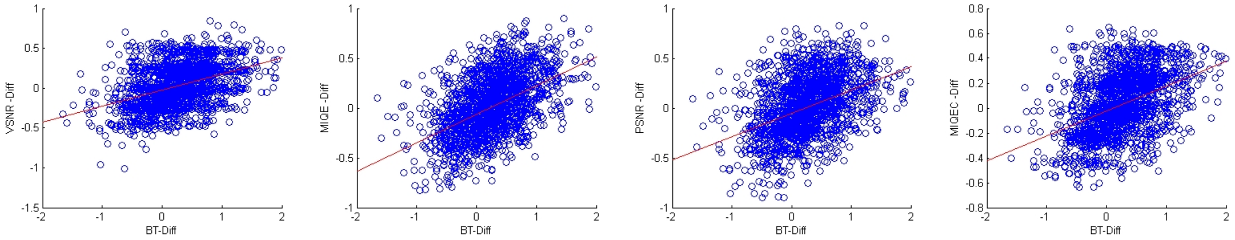


Figure 2.10: Illustration of relationship between QE scores' difference and BT scores' difference when the distortion is JP2K. The QEs and their corresponding Spearman Rank Correlation Coefficients are : a) VSNR (0.399), b) MIQE (0.503), c) PSNR (0.432) and d) MIQEC (0.446)

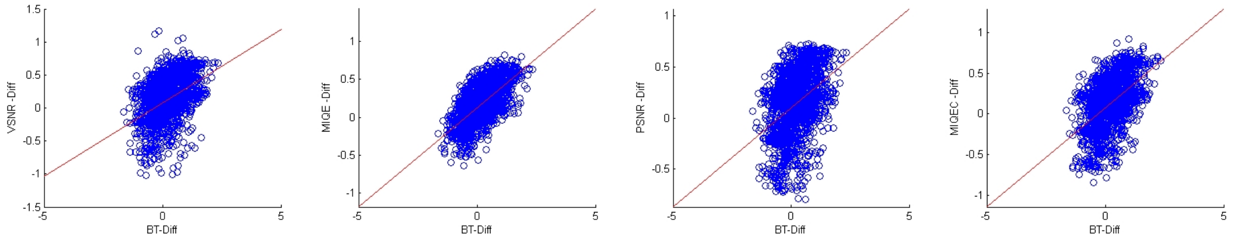


Figure 2.11: Illustration of relationship between QE scores' difference and BT scores' difference when the distortion is JPEG. The QEs and their corresponding Spearman Rank Correlation Coefficients are : a) VSNR (0.464), b) MIQE (0.637), c) PSNR (0.411) and d) MIQEC (0.498)

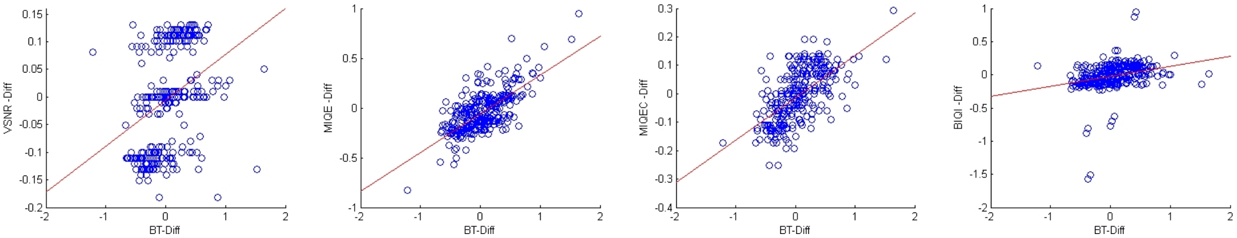


Figure 2.12: Illustration of relationship between QE scores' difference and BT scores' difference when the distortion is Noise. The QEs and their corresponding Spearman Rank Correlation Coefficients are : a) VSNR (0.482), b) MIQE (0.651), c) MIQEC (0.647) and d) BIQI (0.394)

Table 2.7: Correlation Coefficients

	LIVE			TID			CSIQ			Average		
	PCC	SRCC	KRCC	PCC	SRCC	KRCC	PCC	SRCC	KRCC	PCC	SRCC	KRCC
SSIM	0.936	0.939	0.794	0.740	0.775	0.577	0.859	0.876	0.691	0.845	0.863	0.687
VIF	**0.972**	**0.972**	**0.857**	0.777	0.749	0.586	0.925	0.919	0.754	0.891	0.880	0.732
VSNR	0.952	0.939	0.785	0.232	0.705	0.535	0.801	0.811	0.625	0.662	0.818	0.648
MIQE	0.962	0.964	0.838	**0.840**	0.807	0.623	0.916	0.911	0.738	**0.906**	0.894	0.733
PSNR	0.847	0.893	0.727	0.279	0.553	0.403	0.800	0.806	0.608	0.642	0.751	0.579
MSSIM	0.947	0.951	0.818	0.790	0.854	0.657	0.898	0.913	0.739	0.878	0.906	0.738
IWSSIM	0.952	0.960	0.838	0.809	**0.856**	**0.664**	0.903	0.921	0.753	0.888	**0.913**	**0.751**
MIQEC	0.960	0.961	0.828	0.829	0.788	0.608	**0.926**	**0.930**	**0.765**	0.905	0.893	0.734
CWSSIM	0.862	0.894	0.721	0.334	0.528	0.378	0.640	0.674	0.491	0.612	0.699	0.53
BIQI	0.832	0.897	0.720	0.414	0.351	0.244	0.695	0.619	0.442	0.647	0.622	0.469
BRISQUE	0.928	0.941	0.789	0.408	0.322	0.228	0.739	0.556	0.423	0.692	0.606	0.48

We calculate Pearson Correlation Coefficents (PCC), Spearman Rank Correlation Coefficients (SRCC) and Kendall Rank Correlation Coefficients (KRCC) using the corresponding regression approach for each database. The results are provided in Table 2.7. Highlighted values show the highest correlation coefficients. The QE with the highest correlation values changes for different databases and correlation types. However, the correlation coefficients of MIQE and MIQEC are very close to the highlighted values in each case. This is the first subjective test for different resolutions in which these QEs have been compared.

Chapter 3

Quality Estimation for Videos with Different Spatial Resolutions

In this chapter, we propose a video quality estimator for an input video that has different spatial resolution from its associated reference video. Because it estimates quality without resampling the reference or the input videos, it overcomes the drawbacks of QE$_{down}$ and QE$_{up}$. It extends MIQE [62] to consider video by incorporating the effect of motion on the sensitivity of the human vision system (HVS). Further, the proposed quality estimator does not use regression while computing the quality, so it is independent of any subjective dataset. In Section 3.1, we describe our proposed quality estimator, the Multiscale Video Quality Estimator (MVQE). In Section 3.2, we present our subjective test environment, which we created to determine the effectiveness of the proposed QE and compare its performance with other video QEs. In this subjective test, the reference and test videos are displayed at the native resolutions. In Section 3.3, we analyze the results of the subjective test, as well as evaluate the performance of the proposed approach using the Scalable Video Database [6].

3.1 Multiscale Video Quality Estimator

The design of our proposed video quality estimator is based on MIQE [62] to handle the limitations of QE$_{down}$ and QE$_{up}$. However, it also takes into account the temporal characteristics of the video on the perceived quality. Hence, we develop the Multiscale Video Quality Estimator (MVQE) by incorporating the effect of motion information into the MIQE. We describe how we integrate motion into MIQE in detail in the following three subsections. First, we describe how the contrast sensitivity of the HVS depends on the spatial and temporal frequencies. Second, we explain the motion estimation methods. Third, we use the spatio-temporal effects of the HVS to construct our video QE.

3.1.1 The Spatio-Temporal Effect of the HVS on Video Quality

To design a video QE that takes into account the effect of motion and the viewing distance, we should employ the spatio-temporal contrast sensitivity function (STCSF) [43], which describes the sensitivity of the HVS at different spatial and temporal frequencies to changes in visual content. The STCSF has been incorporated into the video QEs described in [63] and [64] which are developed using HVS characteristics.

To incorporate the STCSF, we need to use appropriate frequency units. We use the angular frequency to represent spatial frequency components of two video frames that are viewed with different resolutions and/or different distances. The angular frequency is computed as follows [40]:

$$f(l) = \frac{\pi * d * n}{180 * h * 2^{l+1}} \tag{3.1}$$

In this expression, $f(l)$ denotes the angular (spatial) frequency in cycles per degree (cyc/deg); d, h, and n represent the distance of the viewer, height of the displayed frame, and the number

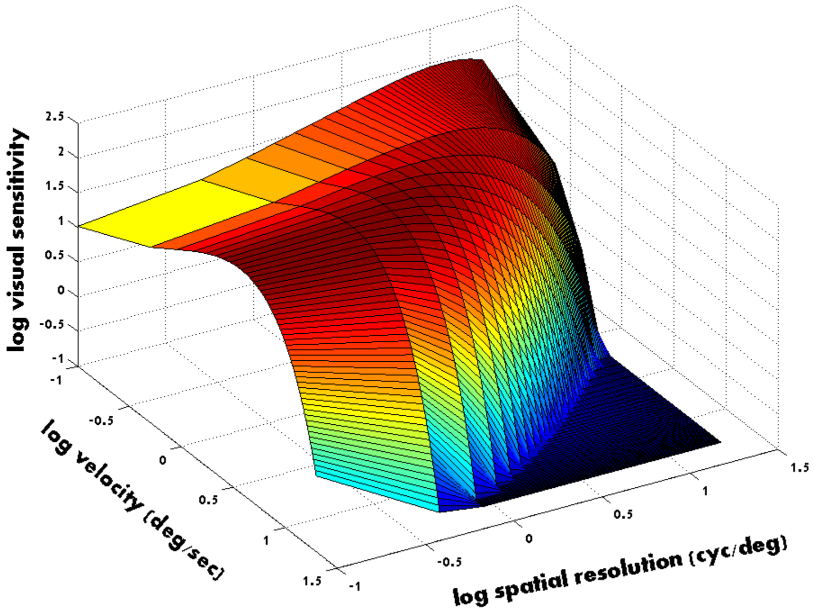


Figure 3.1: Contrast Sensitivity Function

of pixels in the vertical direction of the displayed frame, respectively. Lastly, l indicates the level of a subband decomposition. If there is no subband decomposition, we set $l = 1$. Throughout the text, we denote matrices using bold characters and denote constants and vectors using italics. The temporal frequency is obtained by multiplying the magnitude of the velocity with the spatial frequency [65]. Note that in [62], we had only considered the spatial frequency since there is no need to consider the temporal frequency for images.

The contrast sensitivity of the HVS depends heavily on the spatial and temporal characteristics of the visual content. Kelly [43] measured and modeled the STCSF under stabilized conditions. Daly enhanced this model using unconstrained eye movements as parameters in the model [22]. The contrast sensitivity function of the HVS is illustrated in Figure 3.1. The STCSF has more weight at lower spatial and temporal frequencies, but the spatial resolution which has the maximum sensitivity changes according to the temporal frequency. The sensitivity decreases gradually as either spatial or temporal frequencies increase. As observed in [4] and [5], these characteristics explain why distortions are more tolerable as the size of the visual content decreases and why interpolation of the distorted content makes the artifacts more visible. We use Daly's model [22] to calculate the frequency response of the STCSF. It is computed as follows:

$$\begin{aligned}
&STCSF(f, v_R) = kc_0c_2v_R(c_12\pi f)^2exp\left(-\frac{c_14\pi f}{f_{max}}\right)\\
&k = s_1 + s_2\left|\log(\frac{c_2v_R}{3})\right|^3\\
&f_{max} = \frac{f_1}{c_2v_R + 2}
\end{aligned} \tag{3.2}$$

where the values $s_1 = 6$, $s_2 = 7.3$, $f_1 = 45.9$, $c_0 = 1.14$, $c_1 = 0.67$ and $c_2 = 1.92$ are obtained from [22]. In this equation, f is the spatial frequency in cyc/deg and it is computed using Eq. (3.1). The terms k and f_{max} are utilized to adjust the amplitude of the sensitivity function and the location of the function's peak spatial frequency. The retinal velocity is denoted using v_R and measured in deg/sec. We compute v_R using the following equation [22]:

$$v_R = v_T - v_E \tag{3.3}$$

In this expression, v_T is the velocity of the target object in deg/sec. It depends on the frame rate of the video and spatial frequency and is computed as follows [22]:

$$v_T(f) = \frac{m.FR}{f} \tag{3.4}$$

where, m is the estimated motion of the target object in cyc/frame, FR is the frame rate of the video and f is the spatial frequency. The estimated motion of the target object is computed using one of the estimation methods described in Section 3.1.2. Additionally, v_E denotes the velocity of the eye and it is calculated using the following expression [22]:

$$v_E = \min[(g_{sp}.v_T) + v_{MIN}, v_{MAX}] \tag{3.5}$$

where, $g_{sp} = 0.82$, $v_{MIN} = 0.15$ deg/sec, and $v_{MAX} = 80$ deg/sec. Using Eqs. (3.2)-(3.5), we can find the sensitivity of each spatial frequency and velocity pair. While computing the quality of videos that have different resolutions, we should also take into account the effect of resizing on the computation of the spatial frequency of the visual content. As seen in Eq. (3.1), increasing the viewing distance or decimating by two doubles the spatial frequency of the visual content. Since temporal frequency depends on the motion and spatial frequency, resizing or distance change also indirectly affects the temporal frequency. Algorithms like Video Quality Metric (VQM) [20] and Motion-based Video Integrity Evaluation (MOVIE) [21] take into account temporal characteristics. However, they do not consider the change in frequency-domain representation due to altering the spatial frequency. Hence, they do not fully incorporate spatio-temporal characteristics of the HVS. In addition to using the correct spatial resolution, accurate motion prediction is also necessary to find the sensitivity with high precision.

3.1.2 Motion Estimation

The motion prediction is realized by computing the displacement of objects in space. Here, we only consider the apparent motion and ignore occlusion. Further, the frame rate of the videos that we consider is high enough to prevent flickering. The motion is local in space and frequency. Motion estimation (ME) can be performed between two frames or a group of frames. Different approaches have been proposed including Block Matching, Gradient-Based, and Spatio-temporal Energy-Based methods.

Among these motion estimation approaches, we need to use a method which takes into account the spatial frequency while estimating the motion. Without this, we cannot compute the STCSF value accurately because the STCSF employs retinal velocity that changes according to the spatial frequency as described in Section 3.1.1. We have selected two ME approaches to satisfy this requirement: Hierarchical Block-Based ME using Wavelet (HBME) [66] and Watson-Ahumada Optical Flow-Based Motion Estimation (WAME) [65].

Hierarchical Block-Based ME using Wavelets (HBME)

This approach [66] uses the correlation between the blocks in the consecutive two frames of the video. The first one is called the anchor frame and the second one is called the target frame. These two frames are decomposed into subbands using wavelet-based transforms. The motion is estimated for each subband using block-based motion estimation. HBME also assumes that motion is a correlated among subbands. Hence, the motion information which is found in the lower level is used as a starting point in the next level. The estimation starts with the lowest subband level.

First, the subband of the target frame and anchor frame are divided into non-overlapping 4×4 blocks. Second, a destination block is picked from the target frame's subband. Third, a search range is defined in the anchor frame's subband for the corresponding block. The center of this range collocates with the center of the destination block. Fourth, all the candidate blocks within the search are compared with the destination block. The comparison is performed using the mean absolute difference (l_1 norm). The index of the block which has the minimum difference determines the motion vector for the chosen destination block. These operations are repeated until all the motion vectors are computed. After motion vectors of all blocks are found in the lowest subband level, they are scaled and used as starting locations for the motion estimation in the upper layers.

Watson-Ahumada Optical Flow-Based Motion Estimation (WAME)

This model [65] is based on the optical flow motion estimation. The motion is computed for each point individually and the computation is performed in the frequency domain. The ME

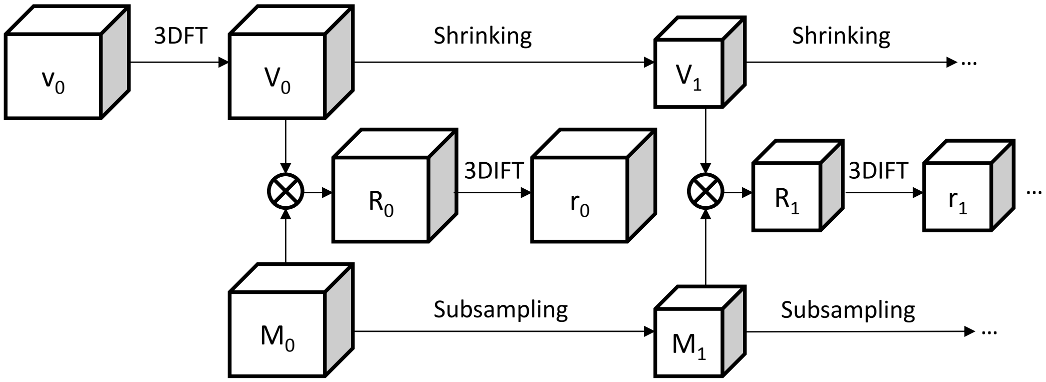


Figure 3.2: Scalar Motion Estimation Block Diagram

algorithm uses the motion trajectory of a particular point and denotes it $c(x, y, t)$. Here, x and y represent the 2D position of the point and t denotes the time index when the point is at location (x, y). Next, the algorithm gets the Fourier transform of this motion trajectory and denotes it $C(u, v, w)$, where u, v and w stand for the spatial frequency components and temporal frequency, respectively. Using these notations, the Fourier transform of the motion is expressed as follows [65]:

$$F(c(x + r_x t, x + r_y t, t)) = C(u, v, w + r_x u + r_y v) \tag{3.6}$$

where $F(.)$ denotes the Fourier transform. The terms r_x and r_y are the motion components in the x and y directions, respectively. According to Eq. (3.6), if the point does not move, the temporal frequency component of the transform will be 0. Otherwise, the amount and the direction of the motion information can be extracted from the shift in the temporal frequency. WAME consists of two stages. In the first stage, *Scalar Motion Computation*, the motion information of a point for different directions and subbands are computed. In the second stage, *Vector Motion Computation*, nonlinear operations are performed to compute the amplitude and the direction of the motion vector for a point at a specific frequency.

These two stages are explained in detail as follows.

In *Scalar Motion Computation* stage, the motion information of a point is found for all frequency and orientation pairs. During implementation, we have used 10 orientations and 4 frequency levels. Figure 3.2 illustrates the block diagram of *Scalar Motion Computation.* We perform scalar motion computation by using the following steps:

1. The video is divided into groups of 8 non-overlapping consecutive frames. We process these video groups in several scales. We denote the video group of the z^{th} scale with $v_z(x, y, t)$ where x, y and t stand for the spatial components and temporal component, respectively. The index of the first scale is 0.

2. The 3D Fourier transform (3DFT) of each group is computed. The Fourier transform of the $v_z(x, y, t)$ is denoted with $V_z(u, v, w)$ where u and v stand for spatial frequencies; w denotes the temporal resolution.

3. A band pass filter group is created which consists of 10 bandpass filters for each scale. If $\vec{s}$ denotes the center spatial frequency of the first band and the orientation index is k (0-9), then $\vec{s}$ can be represented with $(|\vec{s}| * cos(0.2\pi k),\ |\vec{s}| * sin(0.2\pi k))$. The filter for the z^{th} scale is represented with $M_z(\vec{s}, \vec{f}, w)$. The frequency response of the $M_z(\vec{s}, \vec{f}, w)$ is calculated as follows [65]:

$$\begin{aligned}&M_z(\vec{s}, \vec{f}, w) = \{exp[-(\pi\lambda|\vec{s} - \vec{f}|)^2]+\\&exp[-(\pi\lambda|\vec{s} + \vec{f}|)^2]\}[1 - sgn(\vec{s}.\vec{f})sgn(w)]\end{aligned} \tag{3.7}$$

   where $\lambda = 0.795/|\vec{s}|$

4. V_z is filtered using $M_z(\vec{s}, \vec{f}, w)$ and the result is transformed using 3D Inverse Fourier Transform (3DIFT) to obtain the motion at the corresponding frequency and direction.

5. V_{z+1} is obtained by shrinking V_z. The shrinking operation is implemented by filtering

out the frequencies higher than the half of the maximum frequency.

6. $M_{z+1}(\vec{s}, \vec{f}, w)$ is computed by downsampling the filter $M_z(\vec{s}, \vec{f}, w)$.

7. Increment the scale. If the scale is smaller than 4, steps between 1-6 are repeated; otherwise, the procedure is complete.

Vector Motion Computation collects the motion vectors obtained in the previous stage and finds the motion information of a point for a specific frequency. The magnitude of motion vectors at different directions are compared. The motion vector that has the maximum amplitude is selected. The index and the amplitude of this motion vector determine the direction and the magnitude of the estimated motion, respectively. More details about this approach can be found in [65].

3.1.3 Quality Calculation

In this section, we describe how we take into account the motion information to estimate the MVQE. We construct the QE by appropriately weighting the mutual information of each subband block. The calculation of the mutual information for each subband block is similar for image and video. Although the details are explained in [62], for the sake of clarity, we summarize this computation below.

The terms $\mathbf{r}$ and $\mathbf{b}$ represent the reference video frame and the input video frame, respectively. First, we convert pixel values of $\mathbf{r}$ and $\mathbf{b}$ to luminance values using Eq. (2.2) and these values are kept in $\mathbf{r}'$ and $\mathbf{b}'$ matrices. This operation is performed to simulate the adaptation of the pixel values to the HVS.

After that, $\mathbf{r}'$ and $\mathbf{b}'$ are decomposed into their subbands using 9/7 bi-orthogonal wavelet transformation. There are two reasons for using subband decomposition. The wavelet transforms of $\mathbf{r}'$ and $\mathbf{b}'$ are kept in $\mathbf{R}$ and $\mathbf{B}$ matrices, respectively. $\mathbf{R}$ and $\mathbf{B}$ have the same size and they both consist of $3{\times}L+1$ subbands where L denotes the total number of levels. Since the size of $\mathbf{b}'$ is smaller than $\mathbf{r}'$, some of the subbands in $\mathbf{B}$ are missing and we represent these subbands with $\mathbf{0}$ matrices. For instance, if the resampling rate to create the test video is 0.5, the three subbands in the first level of $\mathbf{B}$ are $\mathbf{0}$ matrices. Each subband is represented with its level and orientation indices. Specifically, if $\mathbf{B}_{l,o}$ represents a subband of the $\mathbf{B}$, then l and o stand for the level and orientation, respectively.

Next, we represent $\mathbf{R}$ and $\mathbf{B}$ in terms of distributions. We start by modelling non-overlapping blocks of $\mathbf{R}$'s subbands using Gaussian Scale Mixture (GSM)as in Eq. (2.3).

The index of the block is shown by j, and it can take values from 1 to the number of the blocks in the subband. $\vec{R}_{l,o,j}$ is normally distributed with $N(0, S^2_{l,o,j}\mathbf{C_u})$ where $S_{l,o,j}$ is a random number from a positive scalar random field and $\vec{U}_{l,o,j}$ is a Gaussian vector with mean zero and covariance $\mathbf{C_u}$. Next, blocks of $\mathbf{B_{l,o}}$ are estimated from the blocks of $\mathbf{R_{l,o}}$ using a linear least square estimator (LLSE). We use Eqs. (2.4) and (2.5) to find the attenuation ($A_{l,o,j}$) and additive noise ($\vec{W}_{l,o,j}$) parameters for each $\vec{R}_{l,o,j}$-$\vec{B}_{l,o,j}$ block pair.

Lastly, the *Mutual Information* is calculated for the subband blocks of $\mathbf{R}$ and $\mathbf{B}$. We model $\mathbf{E}$ and $\mathbf{F}$ as being created by passing $\mathbf{R}$ and $\mathbf{B}$ through the HVS, as in [41]. The HVS is modeled as an Additive White Gaussian Noise. First, the mutual information [50] between non-overlapping 4x4 blocks of $\mathbf{E_{l,o}}$ and $\mathbf{R_{l,o}}$ ($I(\mathbf{E_{l,o,j}}; \mathbf{R_{l,o,j}})$) is computed as follows:

$$I(\mathbf{E_{l,o,j}}; \mathbf{R_{l,o,j}}|s_{l,o,j}) = \frac{1}{2}log_2\left(\frac{s^2_{l,o,j}\mathbf{C_u} + \sigma^2_n\mathbf{I}}{\sigma^2_n\mathbf{I}}\right) \tag{3.8}$$

where $s_{l,o,j}$ is a realization of $S_{l,o,j}$. If $s_{l,o,j}$ is known, $\vec{R}_{l,o,j}$ has a Gaussian distribution.

Second, the mutual information between the blocks of $\mathbf{F_{l,o}}$ and $\mathbf{R_{l,o}}$ is found.

$$I(\mathbf{F_{l,o,j}};\mathbf{R_{l,o,j}}|s_{l,o,j}) = \frac{1}{2}log_2\left(\frac{s^2_{l,o,j}A^2_{l,o,j}\mathbf{C_u}+\sigma^2_{q_{l,o,j}}\mathbf{I}}{\sigma^2_{q_{l,o,j}}\mathbf{I}}\right) \tag{3.9}$$

where, $\sigma^2_{q_{l,o,j}} = \sigma^2_{w_{l,o,j}} + \sigma^2_n$. We use Eqs. (3.8) and (3.9) to find the mutual information for each subband.

Next, we scale ($I(\mathbf{E_{l,o,j}};\mathbf{R_{l,o,j}})$, $I(\mathbf{F_{l,o,j}};\mathbf{R_{l,o,j}})$) according to the size and the importance of the subband block. To calculate the magnitude of scaling coefficients, we take into account two factors: the size of each subband, and the visual importance of each subband block determined by the characteristics of the HVS. In MIQE, the weighting coefficients only depend on the first factor. Our video QE here incorporates the impact of motion in the STCSF as described in Eq. (3.2). Since the motion may vary according to the position of the subband, the weight of each subband block is computed separately in MVQE.

First, the HVS has a different weight for each subband block. As expressed in Eq. (3.2), the STCSF depends on the motion and the spatial frequency. Since the motion information is local and varies with spatial frequency, the value of the contrast sensitivity can change for each block.

Second, the effect of subband size should be compensated. As the subband level increases by one, the size of the band decreases by four. STCSF only takes into account the influence of the spatial and temporal frequencies on the perception of visual quality. It does not take into account the effect of the number of samples in each frequency band. Rouse et al. [52] show that incorporating this improves the accuracy of quality estimation. Therefore, to compensate for the size of each level l, the mutual information of each subband block is scaled by 2^{2l}. Hence, the scaling coefficient of each subband block is computed using the following expression:

$$G_{l,o,j} = STCSF(f_{l,o}, v_{R,l,o,j}) * 2^{2l} \tag{3.10}$$

where, $f_{l,o}$ is the spatial frequency for subband (l, o) and $v_{R,l,o,j}$ is the retinal velocity for subband block (l, o, j).

Finally, the similarity of each subband block is scaled with the corresponding weight and the value of the estimated quality for a particular frame is found using the following formula:

$$FQ = \frac{\sum_{l,o,j\ \in\ subband\ blocks} G_{l,o,j} * I_{B,l,o,j}}{\sum_{l,o,j\ \in\ subband\ blocks} G_{l,o,j} * I_{R,l,o,j}} \tag{3.11}$$

In this expression, FQ denotes the quality of the frame, $G_{l,o,j}$ is the scaling coefficient for block (l, o, j) given in Eq. (3.10). $I_{B,l,o,j}$ and $I_{R,l,o,j}$ represent the mutual information of the input and the reference signals for the subband block (l, o, j). FQ takes into account the joint effect of motion and spatial resolution on the perception using STCSF. However, it denotes the quality of a particular frame. Hence, we average FQ values for all frames in the video to obtain the video quality as follows:

$$MVQE = \frac{1}{N} * \sum_{k=1}^{N} FQ_k \tag{3.12}$$

The final quality value is between 0 and 1, and this value increases as the quality increases.

To check the validity of the proposed algorithm, it is necessary to compare the estimated quality values with the preference of viewers. In the following section, we describe the details of our subjective tests to obtain viewer preferences.

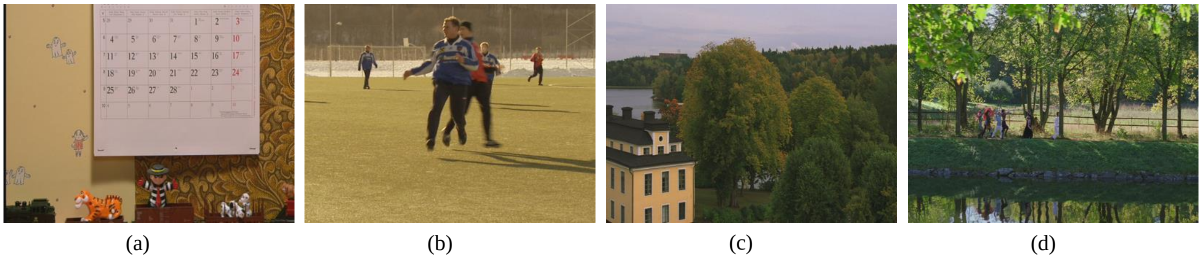


Figure 3.3: Snapshots of subjective test videos : (a) Mobile Calender (b) Soccer (c) IntoTree (d) ParkJoy

3.2 Subjective Tests

We perform subjective tests to obtain the viewers' mean opinion scores of videos that have different spatial resolutions. The test results are used to ensure that the proposed algorithm works properly. They also provide a benchmark to compare our algorithm with existing approaches. Our subjective tests are performed by displaying a group of videos with different compression and resolution parameters at their native resolutions to the viewers to obtain their scores. In the following subsections, we describe how we create the test set, perform subjective tests, and analyze results.

3.2.1 Test Set Creation

There are in total 4 reference videos with different spatial and temporal complexities. They are chosen from a public database in [67] where the original sources are referenced. The name of the videos are *Soccer*, *MobileCalendar*, *IntoTree* and *ParkJoy*. *MobileCalendar*, *IntoTree* and *ParkJoy* have a 1280 × 720 spatial resolution; their frame rate is 50 fps. On the other hand, the spatial resolution of *Soccer* is 704×576 and its frame rate is 60 fps. The snapshots of the videos are illustrated in Figure 3.3. The videos have different spatial and temporal characteristics. To quantify the spatial and temporal complexities of the videos, we used

Table 3.1: SI and TI values of the test videos

	MobileCalendar	Soccer	TreeInto	ParkJoy
SI	106.165	77.641	68.196	126.477
TI	23.462	37.444	16.471	37.837

spatial information (SI) and temporal information (TI) indices as described in ITU-R 910 [68]. The SI and TI values of the videos are shown in Table 3.1.

We have created nine test videos for each reference video. Of these nine test videos, five are high-resolution and are created by compressing a reference video. The remaining four videos are low-resolution test videos. These videos are obtained by decimating the reference video by a factor of 0.5 and compressing the decimated video. We compress the full-size reference video and the decimated reference video using the H.264 AVC JM 18.1 [69] reference software. We employ the Non-Normative Filter as a low pass filter during decimation. It is a Sine-windowed Sinc-function and is formulated in [70]:

$$f(x) = \begin{cases} \frac{sin\left(\frac{\pi * X}{D}\right)}{\frac{\pi * X}{D}} sin\left(\frac{\pi}{2}(1 + \frac{x}{N * D})\right) & |x| < N * D \\ 0 & \text{Otherwise} \end{cases} \tag{3.13}$$

where D is the decimation parameter and N represents the number of lobes. The parameters are $D = 2.5$ and $N = 3$ and phase offset$= 0$, and the implementation is performed using the *DownConvert* application of JSVM codec. The corresponding filter is the default mode of the application.

Test videos should cover a range from excellent quality to very bad quality. We utilize the quantization parameter instead of the bitrate to select the test videos because the spatial and temporal complexities of the test videos are dramatically different. We should also observe the effect of the bitrate on the quality of high and low resolution test videos at different quality regions. Therefore, QP values of the low resolution test videos are determined according to the QP values of the high resolution test videos. We use the following notations to

Table 3.2: QPs and Bitrates (BRs) of Test Videos

	MobileCalendar			Soccer			IntoTree			ParkJoy		
Index	QP	HR BR	LR BR	QP	HR BR	LR BR	QP	HR BR	LR BR	QP	HR BR	LR BR
1	28	1166.9	295.4	28	599	175.3	28	1156	235.9	28	4836.9	901.9
2	31	724.4	202.6	31	381.6	118.2	31	664.1	144.1	32	2766.8	502.9
3	37	296.9	86.7	36	172.4	56.1	36	233.2	55.6	39	868.4	159.2
4	42	139.1	39.6	39	113.6	37.8	39	135.1	33.5	42	463.5	87.2
5	45	89.1		44	54.7		44	55.3		47	149.9	

describe the relationships between the bitrate and the quantization values of high-resolution and low-resolution test videos. HR and LR denote high-resolution and low-resolution test videos, respectively. The index of the test video for a specific resolution is denoted by β. Hence, QP_{β}^{HR} represents the QP value of high-resolution test video at the β^{th} index, and BR_{β}^{HR} denotes the bitrate of this video. The QP value increases (quality decreases) as β increases. Based on this notation, we choose the bitrates of the test videos as follows. First, we assume that $QP_{1}^{HR} = 28$ and $QP_{1}^{LR} = 28$ for all reference video types.

If β is even ($\beta \in 2, 4$), then BR_{β}^{HR} will be found by averaging $BR_{\beta-1}^{HR}$ and $BR_{\beta-1}^{LR}$. QP_{β}^{HR} and QP_{β}^{LR} are equal to the QP value of the high-resolution test video which has the closest bitrate to the BR_{β}^{HR}. On the other hand, if β is odd ($\beta \in 3, 5$), we set BR_{β}^{HR} to $BR_{\beta-2}^{LR}$ and the QP value which satisfies this requirement is assigned to QP_{β}^{HR}. The bitrates and the QP values of the test videos are in Table 3.2. The unit of the bitrate is $KBps$. More information about the video database can be found at [71].

To understand the joint effect of quantization and resolution clearly, we should design the subjective tests carefully. The subjective test results should give us score estimates. To have reliable quality estimates [72], the viewers should playback the videos anytime during the test. This can be achieved using the Subjective Assessment Methodology for Video Quality (SAMVIQ) [73] method. This method allows the viewer to access any video more than once. A viewer can also update the score based on their evaluation of other videos. During these updates, the viewer implicitly ranks videos. More details about the implementation of the test is given in the next subsection.

Table 3.3: Mean Opinion Scores and ANOVA F Scores

	H1	L1	H2	L2	H3	L3	H4	L4	H5	F	p
MobileCalendar	88.9	74.2	83.5	70.4	63.7	48.2	43.0	30.8	31.7	62.53	$4.55*10^{-52}$
Soccer	90.5	73.8	86.7	64.2	70.7	53.0	58.2	42.5	36.9	37.87	$3.47*10^{-37}$
IntoTree	88.8	71.7	82.4	62.0	60.8	42.7	42.7	28.8	20.8	58.13	$9.78*10^{-50}$
ParkJoy	89.1	72.8	85.0	66.2	65.2	43.0	48.6	27.6	24.6	56.33	$9.60*10^{-49}$

Note that using more conventional approaches can cause several drawbacks. Single Stimulus (SS) and Double Stimulus (DS) [74] based methods are effective when the reference video and the test videos have equal spatial resolutions. However, characterizing the qualities of videos that have different spatial resolutions is more challenging. First, the videos have multiple distortions and viewers may have difficulties scoring videos by watching them once. Second, video scores may depend on the order in which they are presented, i.e. the quality of the videos which are at the beginning of the test could be underestimated or overestimated because the remaining videos have not been viewed yet. Third, videos seen earlier cannot be reviewed later in the test; so it is not possible to exploit a viewed video while determining the quality of the remaining videos. Fourth, as explained in [6], when the quality of the test videos are similar or hard to score, comparison based approaches are more reliable than absolute category rating approaches like DS or SS. On the other hand, the paired comparison approach can provide a reliable ranking among different videos. Nevertheless, the paired comparison method can also be improved by using the actual opinion scores while ranking the test videos.

3.2.2 Subjective Test Implementation

The settings for our subjective test environment were based on the requirements of ITU-R 910 [68]. The tests were performed in a low-illuminated room. The color of the wall behind the monitor was gray. We padded the sequences with mid-gray to cover the full screen. The color temperature of the monitor was 6500 K°. The test contained a total of 40 videos, and 25 people performed the tests. Each test session took approximately 40 minutes. A training session is performed before the testing session to familiarize the viewers with the test. The viewers were either graduate or undergraduate students. They had clear vision and they were non-experts. The test session consisted of 4 groups. In each group, a reference video and its corresponding 9 test videos were displayed to the viewers. There were 4 low-resolution and 5 high-resolution test videos for each group. Each group starts with the corresponding reference video and by pressing 0 the reference video could be displayed again as described in [73]. However, the indices of the test videos changed randomly for each test session and each test group to prevent bias. The order of test groups was also random.

We asked the viewers to give a score for each test video between 0 and 100. The score of the reference video was 100. The SAMVIQ [73] methodology instructs the viewers that: Bad (0-20), Poor (21-40), Fair (41-60), Good (61-80), Excellent (81-100). Each video was 10 seconds long. Viewers were encouraged to watch the whole video a few times before finalizing their decisions. The distance between the viewers and the screen was 6 times the height of the 640 × 360 frame. Tests were performed using Dell XPS 8100 with a 12 GB RAM and NVidia Geforce GT640 graphics card. The display was a Dell 2212HM with a resolution of 1920 × 1080. Hence, the values of n, h and d in Eq. (3.1) are 360 pixels, 89.1 mm and approximately 6×h, respectively, for the low resolution test video and they are 720 pixels, 178.2 mm and approximately 3×h, respectively, for the high resolution test video.

3.3 Performance Evaluation

In this section, we compare the performance of our proposed video QE to 13 approaches using the results of the subjective test. We begin by evaluating the statistical characteristics of the subjective test. Then, we examine the similarity between subjective test scores and QE scores using QE_{down} and QE_{up} setups. Next, we explore similarity scores of QEs for two specific quality regions. As a final step, we evaluate the performance of MVQE using the

Scalable Video Database [6].

We have used the method described in Rec. ITU-R BT.500-11 [74] to perform consistency check on the subjective test scores. According to the results of this check, we do not have any outlier in our subjective test. We have also employed Analysis of Variance (ANOVA) [75] to assess whether the distribution of subjective scores for the test videos in the same video group is different. We have used *anova1* function of Matlab to compute the F scores and p values of ANOVA analysis. According to ANOVA, F_{crit} value is 1.98 for 95% confidence interval. Table 3.3 shows the MOSs, F scores and p values for subjective test results. According to Table 3.3, all F scores are greater than F_{crit} ($p < 0.05$), therefore the subjective scores satisfy the 95% confidence interval. Next, we calculate the similarity between MOSs and QEs using the following correlation metrics: Pearson, Spearman Rank Correlation Coefficient (SRCC), Kendall Rank Correlation Coefficient (KRCC), Root Mean Square Error (RMSE) and Outlier Ratio (OR). These metrics [76] measure the ranking similarity, accuracy, linearity and consistency of QEs. We also compute the Fraction of Correct Ranking (FCR) and Fraction of Correct Similarity (FCS) [62] metrics. Pearson, SRCC, KRCC, RMSE and OR consider the relationship between all test video scores, but FCR and FCS only examine the relationship between HR and LR test videos. In addition to measure these metrics, we check whether the QEs are statistically different for a specific correlation metric. To test the statistical difference of two QEs' SRCC scores, we use the following formula [77]:

$$\delta(\eta - 3) = \frac{(\rho_{13} - \rho_{23}) * \sqrt{(1 + \rho_{12}) * (\eta - 1)}}{\sqrt{2 * \kappa * \frac{\eta-1}{\eta-3} + \frac{(\rho_{13}+\rho_{23})^2}{4} * (1 - \rho_{12})^3}} \tag{3.14}$$

where δ and η denote the statistical difference and number of test videos, respectively. SRCC scores of the QEs are denoted with ρ_{13} and ρ_{23}, respectively. We also compute the spearman correlation coefficient between the QEs and denote it with ρ_{12}. Lastly, κ is computed as follows:

$$\kappa = 1 - \rho_{13}^2 - \rho_{23}^2 - \rho_{12}^2 + 2 * \rho_{13} * \rho_{23} * \rho_{12} \tag{3.15}$$

If η is greater than or equal to 30, δ has to be greater than 1.96 to have statistically different metrics [77]. If η is smaller than 30, we use the t-distribution table to find the threshold level and check whether δ is greater than this value. While analyzing the results, we check whether the QE metric with the maximum SRCC score is statistically different from other QEs. We indicate statistically indistinguishable QEs' SRCC scores with " * " in the tables.

We use both QE_{down} and QE_{up} setups to compute the QE values for the full-reference QEs that are designed for visual contents with same spatial resolutions. During QE_{down}, we use the Non-Normative filter as the decimation filter, while we used the Lanczos 3 filtering after upsampling in QE_{up}. We perform the filtering via DownConvert application in JSVM. Compared QE methods that have to employ QE_{down} or QE_{up} are PSNR, VQM [20], MOVIE [21], MIQE [62], SSIM [18], MSSIM [19]. In addition to these FR QEs, we use scalable QEs [26], [27], [28] in our comparisons. We denote these estimators with SQM1, SQM2 and SQM3, respectively. Note that, SQM1, SQM2 and SQM3 are developed using regression parameters and we use the parameters given in the corresponding papers. In addition to that, we resize the videos to compute SQM2 and SQM3 whenever it is required. We also measure the performance of three NR QEs: NR-BLUR [14], NR-GBIM [15] and NR-BIQI [16]. NR-BLUR and NR-GBIM compute the quality of the video frame using the amount of blur and blocking, respectively. On the other hand, NR-BIQI is a two-step framework and computes the quality using natural scene statistics. In addition to these metrics, we also test the performance of MVQE without using G, STCSF or 2^{2l} in Eq.(3.10) to understand the influence of scaling coefficients. We denote these metrics with MVQE-noG, MVQE-noS and MVQE-noL, respectively. We use logistic regression as described in [78] to remove the nonlinearity of QE scores. The logistic regression function is as follows:

Table 3.4: Metric Scores of QEs

	Pearson	SRCC	KRCC	RMSE	OR	FCR	FCS
PSNR-QE_{down}	0.851	0.836	0.648	10.937	0.444	0.788	0.559
VQM-QE_{down}	0.937	0.927	0.756	7.293	0.278	0.850	0.677
MOVIE-QE_{down}	0.930	0.938	0.803	7.676	0.306	0.875	0.706
SSIM-QE_{down}	0.856	0.840	0.663	10.752	0.500	0.763	0.530
MSSIM-QE_{down}	0.833	0.804	0.619	11.515	0.667	0.700	0.449
PSNR-QE_{up}	0.778	0.709	0.546	13.101	0.444	0.763	0.510
VQM-QE_{up}	0.974	0.968*	0.844	4.693	**0.056**	0.925	0.788
MOVIE-QE_{up}	0.912	0.907	0.749	8.566	0.389	0.788	0.572
SSIM-QE_{up}	0.851	0.823	0.638	10.951	0.417	0.775	0.545
MSSIM-QE_{up}	0.922	0.914	0.765	8.074	0.194	0.862	0.667
MIQE	0.938	0.943	0.825	7.213	0.167	0.938	0.815
MVQE-HB	0.973	0.973*	**0.883**	4.782	0.083	**0.975**	**0.880**
MVQE-WA	**0.976**	**0.977**	0.879	**4.583**	**0.056**	0.963	0.860
MVQE-noG	0.742	0.739	0.543	14.100	0.611	0.675	0.370
MVQE-noS	0.868	0.864	0.682	10.327	0.444	0.825	0.627
MVQE-noL	0.925	0.928	0.781	7.901	0.278	0.888	0.727
SQM1	0.908	0.925	0.769	8.717	0.333	0.900	0.748
SQM2	0.869	0.860	0.679	10.318	0.444	0.800	0.575
SQM3	0.237	0.494	0.356	20.239	0.750	0.600	0.279
NR-BIQI	0.524	0.519	0.387	17.740	0.639	0.725	0.481
NR-BLUR	0.612	0.486	0.381	16.471	0.583	0.663	0.423
NR-GBIM	0.857	0.860	0.670	10.736	0.417	0.850	0.665

$$LR(Q) = \frac{b_1 - b_2}{1 + e^{\frac{-(Q-b_3)}{|b_4|}}} + b_2 \tag{3.16}$$

where Q denotes the estimated quality score and b_1, b_2, b_3 and b_4 are the parameters of the logistic regression function. Table 3.4 shows the correlation scores between QEs and MOSs for all videos. The columns of the table represent the metrics and the rows of Table 3.4 show different QE approaches. We add QE_{down} and QE_{up} suffixes to the QEs to denote which resizing approach is used during QE computation. WA and HB suffixes stand for Watson-Ahumada optical flow based and Hierarchical Block-based motion estimation models, respectively. According to the table, SQM3 has the lowest scores. MVQE-HB and MVQE-WA algorithms perform better than the other QEs. MIQE's correlation scores are lower than the proposed approaches. Hence, integrating motion information has improved the QE estimation. MVQE-WA also outperforms MVQE-noG and MVQE-noS. MVQE-noG performs even worse than PSNR. Therefore, we can infer that employing scaling coefficients

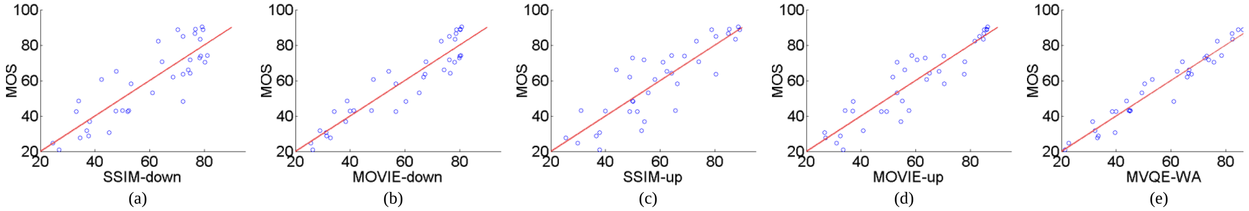


Figure 3.4: Illustration of relationship between MOS scores and QE Scores for the following quality models : (a) SSIM-QE$_{down}$, (b) MOVIE-QE$_{down}$, (c) SSIM-QE$_{up}$, (d) MOVIE-QE$_{up}$, (e) MVQE-WA

provides a noticeable improvement although this improvement is not as big as the contribution of mutual information scores. We have also observed that the correlation scores of MVQE-noL are higher than the correlation scores of MVQE-noS. Hence, the improvement due to STCSF is more than the contribution of Subband Size Compensator (2^{2l}). VQM-QE_{up} has slightly lower scores than the proposed approaches for Pearson and SRCC metrics and OR scores of MVQE-WA and VQM are the same. The difference increases when we look at the KRCC and RMSE scores. FCR and FCS scores of the proposed approaches are also higher than others. Both NR metrics and SQMs have lower correlation scores than the proposed approaches. MVQE-WA is statistically different from all the metrics except VQM-QE_{up}. We also illustrate the relationship between MOS and several QEs in Figure 3.4, for SSIM-QE$_{down}$, MOVIE-QE$_{down}$, SSIM-QE$_{up}$, MOVIE-QE$_{up}$ and MVQE-WA.

To obtain more insight about the robustness of the QE scores, we should also examine the correlation metrics in different quality regions. As explained in Section 2.1, both QE_{down} and QE_{up} have several limitations. However, the effects of these limitations might not be apparent in certain conditions and may become more problematic when these conditions change. To explore the efficiency and the consistency of each approach, we divide the test videos into 2 regions. Test videos with a MOS greater than 50 are collected into the High-Quality Region. The remaining videos are put into the Low-Quality Region.

In Table 3.5, we demonstrate the results of similarity scores for the videos that are in the

Table 3.5: Metric Scores for Low Quality Region Videos

	Pearson	SRCC	KRCC	RMSE	OR	FCR	FCS
PSNR-QE_{down}	0.663	0.506	0.383	7.744	0.625	0.688	0.363
VQM-QE_{down}	0.860	0.776	0.583	5.272	0.500	0.750	0.475
MOVIE-QE_{down}	0.895	0.847*	0.650	4.612	0.313	0.750	0.475
SSIM-QE_{down}	0.731	0.709	0.583	7.058	0.500	0.688	0.388
MSSIM-QE_{down}	0.679	0.632	0.483	7.590	0.563	0.625	0.325
PSNR-QE_{up}	0.634	0.434	0.305	7.998	0.688	0.563	0.338
VQM-QE_{up}	0.847	0.800	0.583	5.491	0.438	0.750	0.475
MOVIE-QE_{up}	0.753	0.759	0.533	6.869	0.625	0.563	0.263
SSIM-QE_{up}	0.685	0.665	0.467	7.534	0.688	0.625	0.325
MSSIM-QE_{up}	0.821	0.806	0.583	5.904	0.313	0.750	0.450
MIQE	0.770	0.762	0.617	6.593	0.438	0.813	0.563
MVQE-HB	**0.907**	**0.941**	**0.833**	**4.351**	**0.250**	**1.000**	**0.775**
MVQE-WA	0.894	0.891*	0.733	4.624	0.313	0.875	0.625
MVQE-noG	0.289	0.271	0.167	9.900	0.750	0.438	0.162
MVQE-noS	0.622	0.644	0.542	8.097	0.625	0.562	0.238
MVQE-noL	0.745	0.794	0.633	6.902	0.625	0.875	0.625
SQM1	0.707	0.702	0.518	7.310	0.563	0.813	0.538
SQM2	0.664	0.521	0.400	7.736	0.625	0.688	0.363
SQM3	0.286	0.278	0.234	9.954	0.813	0.500	0.250
NR-BIQI	0.419	0.456	0.317	9.390	0.563	0.625	0.325
NR-BLUR	0.192	0.176	0.133	10.148	0.563	0.500	0.225
NR-GBIM	0.453	0.476	0.317	9.218	0.688	0.750	0.450

Table 3.6: Metric Scores for High-Quality Region Videos

	Pearson	SRCC	KRCC	RMSE	OR	FCR	FCS
PSNR-QE_{down}	0.516	0.469	0.384	8.616	0.750	0.563	0.238
VQM-QE_{down}	0.694	0.741	0.516	7.242	0.750	0.708	0.458
MOVIE-QE_{down}	0.684	0.788	0.653	7.355	0.800	0.750	0.488
SSIM-QE_{down}	0.451	0.463	0.347	8.981	0.850	0.500	0.226
MSSIM-QE_{down}	0.387	0.313	0.263	9.279	0.850	0.375	0.125
PSNR-QE_{up}	0.860	0.735	0.547	5.131	0.400	0.708	0.411
VQM-QE_{up}	**0.957**	0.920*	0.758	**2.914**	**0.200**	0.917	0.702
MOVIE-QE_{up}	0.926	0.788	0.653	3.791	0.300	0.708	0.399
SSIM-QE_{up}	0.803	0.735	0.547	5.994	0.500	0.708	0.399
MSSIM-QE_{up}	0.814	0.729	0.600	5.850	0.350	0.708	0.399
MIQE	0.949	0.890*	0.747	3.187	0.200	0.917	0.702
MVQE-HB	0.936	0.880*	0.737	3.546	0.150	0.917	0.702
MVQE-WA	0.955	**0.925**	**0.779**	2.970	**0.200**	**0.958**	**0.768**
MVQE-noG	0.919	0.841	0.693	3.964	0.300	0.708	0.400
MVQE-noS	0.912	0.779	0.612	4.123	0.350	0.708	0.399
MVQE-noL	0.890	0.796	0.614	4.710	0.400	0.750	0.440
SQM1	0.592	0.779	0.583	8.109	0.900	0.792	0.530
SQM2	0.577	0.592	0.505	8.219	0.700	0.667	0.387
SQM3	0.502	0.614	0.421	8.704	0.650	0.708	0.411
NR-BIQI	0.310	0.244	0.203	9.565	0.700	0.542	0.446
NR-BLUR	0.390	0.410	0.343	9.263	0.800	0.625	0.357
NR-GBIM	0.562	0.535	0.379	8.322	0.600	0.708	0.446

Low-Quality Region. There is a slight decrease in MVQE-HB and MVQE-WA correlation metric scores with respect to the scores in Table 3.4. Nevertheless, they still have the highest metric scores. MOVIE-QE_{down} has the third highest score among all QEs. The decrease in QE_{up}-based approaches are much higher than the decrease in the proposed approaches. For instance, in Table 3.4, the Kendall metric difference between MVQE-WA and VQM-QE_{up} is 0.035, but this difference is 0.15 in Table 3.5. This difference change occurs because of QE_{up}'s limitations. As described in Section 2.1, QE_{up}-based approaches underestimate the quality of low-resolution test videos. However, people prefer LR test videos more than HR test videos in the low quality region.

We introduce the following definitions to clarify this discussion. HR_{winner} denotes the total number of times the MOS of a HR test video is higher than the MOS of a LR test video, and LR_{winner} stands for the total number of cases where the MOS of a LR test video is higher than that of a HR test video. During the computation of HR_{winner} and LR_{winner}, comparisons between HR and LR videos are performed for each content separately. In the Low-Quality Region, LR_{winner}, 9, is higher than HR_{winner}, 7. Therefore, we observe the rank mismatch between MOSs and QE_{up}-based approaches in Table 3.5. In the low quality region, MVQE-WA is statistically different from VQM-QE_{up} but it is not statistically different from MOVIE-QE_{down}.

Table 3.6 shows the correlation scores when the test videos are in the High-Quality Region. According to the table, MVQE-WA has the highest scores for SRCC, KRCC, FCR and FCS metrics. However, VQM-QE_{up} has the highest Pearson, RMSE and OR scores. NR-BLUR has the lowest scores. Moreover, the correlation scores of QE_{down}-based approaches are lower than those of the QE_{up}-based approaches. This is due to the characteristics of QE_{down} QEs. QE_{down} QEs overestimate the quality of low-resolution videos. This limitation can be observed if HR_{winner} is higher than LR_{winner}. For instance, for the set of the test videos in the High-Quality Region, HR_{winner} is 17 and LR_{winner} is 7. As in Table 3.4, MVQE-WA is

Table 3.7: Metric Scores for H.264 SVC Coded Videos

	Pearson	SRCC	KRCC	RMSE	OR
PSNR	0.768	0.795	0.609	12.244	0.286
VQM	0.851	0.746	0.556	10.026	0.357
MOVIE	0.913	0.838	0.646	7.808	0.179
SSIM	0.907	0.872	0.672	8.060	0.250
MSSIM	0.851	0.882	0.688	10.035	0.321
MIQE	0.935	**0.927**	**0.767**	6.809	**0.107**
MVQE-HB	0.912	0.897*	0.720	7.841	0.250
MVQE-WA	**0.937**	0.924*	**0.767**	**6.711**	**0.107**
MVQE-noG	0.929	0.897*	0.725	7.033	**0.107**
MVQE-noS	0.927	0.907*	0.730	7.160	0.178
MVQE-noL	0.916	0.872	0.693	7.669	0.286
SQM2	0.770	0.808	0.630	12.193	0.286
SQM3	0.557	0.715	0.561	15.874	0.536
SQM4	0.874	0.884*	0.698	9.297	0.357
NR-BIQI	0.575	0.566	0.470	15.636	0.536
NR-BLUR	0.748	0.725	0.534	12.689	0.500
NR-GBIM	0.094	0.218	0.138	19.029	0.643

statistically different from all the metrics except VQM-QE_{up}.

MVQEs have been designed to compare videos with different spatial resolutions at their native resolutions. However, we should also consider the case when the test video is spatially and temporally interpolated at the end user's device. For instance, we can perform optimum video transmission using scalable video coding. In this case, the reference and modified test videos can have same spatial and temporal resolutions. To evaluate the performance of MVQE-WA and MVQE-HB in this scenario, we can utilize Scalable Video Database [6]. This database consists of test videos that are encoded either using H.264 Scalable Video Coder (SVC) or Wavelet SVC (WSVC). These videos are created using a combination of spatial, temporal and SNR scalability parameters. Before subjective tests, they are spatially or temporally upsampled to the spatial and temporal resolution of the reference video. We calculate correlation scores of the QEs after applying logistic regression approach in Eq. (3.16) to the QEs. We add Feghali's QE [23] to the QEs and denote it with SQM4. SQM4 is computed by adding PSNR with a function of temporal resolution and motion activity. Therefore, the performance of SQM4 and PSNR is the same if there is no change in the temporal resolution. However, in the Scalable Video Database, some videos are encoded

Table 3.8: Metric Scores for WSVC Coded Videos

	Pearson	SRCC	KRCC	RMSE	OR
PSNR	0.735	0.754	0.565	13.989	0.405
VQM	0.834	0.786	0.595	11.366	0.262
MOVIE	0.843	0.757	0.586	11.067	0.310
SSIM	**0.907**	0.883*	**0.723**	**8.656**	**0.190**
MSSIM	0.899	0.842*	0.679	9.003	0.262
MIQE	0.863	0.852*	0.665	10.402	0.357
MVQE-HB	0.873	0.861*	0.686	10.020	0.214
MVQE-WA	0.882	0.886*	0.702	9.709	0.214
MVQE-noG	0.872	0.854*	0.665	10.291	0.286
MVQE-noS	0.898	**0.890**	0.711	9.611	0.238
MVQE-noL	0.881	0.856*	0.665	9.742	0.286
SQM2	0.749	0.745	0.562	13.640	0.333
SQM3	0.697	0.831*	0.648	14.749	0.476
SQM4	0.798	0.780	0.600	12.404	0.405
NR-BIQI	0.101	0.070	0.058	20.476	0.619
NR-BLUR	0.545	0.561	0.383	17.256	0.548
NR-GBIM	0.611	0.594	0.491	16.293	0.548

using temporal scalability. Note that, we cannot compute SQM1 for this database because the quantization parameters are not available for these videos. Tables 3.7 and 3.8 provide the metric scores for H.264 SVC and WSVC videos, respectively. The correlation scores of MIQE and MVQE-WA are higher than other approaches when the videos are encoded using H.264 SVC. On the other hand, SSIM and MVQE-WA outperform other approaches if the test videos are encoded using WSVC. Although MVQE-WA is not statistically different from VQM-QE_{up} for some of the results in Tables 3.4 and 3.6, MVQE-WA is statistically different from VQM for all test videos in the Scalable Video Database. Moreover, according to Tables 3.7 and 3.8, MVQE-WA performs slightly better than MVQE-noS, MVQE-noG and MVQE-noL. Hence, we can infer that using scaling coefficients slightly improves the performance of MVQE when the quality estimation is performed for different resolution videos that are displayed at the same resolution. We also compute the running time of the QEs. Table 3.9 reports the relative running time of each algorithm as computed on the same computer used for the subjective tests, where all run times are relative to the time needed to compute PSNR.

Table 3.9: Relative Running Time of QEs to the Running time of PSNR

QE	Comp	QE	Comp	QE	Comp
PSNR	1	MIQE	165	SQM3	738.7
VQM	56.5	MVQE-HB	1495.4	SQM4	669.2
MOVIE	8002.7	MVQE-WA	4741.8	NR-BIQI	89.7
SSIM	21.7	SQM1	1	NR-BLUR	129
MSSIM	40.1	SQM2	666.2	NR-GBIM	3073.3

Chapter 4

Reduced Complexity Perceptually Optimum Multicasting

In this chapter, we work on problem of optimum multicasting of videos to the devices with different spatial resolutions using perceptual QEs. We specifically focus on the complexity of the optimum perceptual multicasting that increases due to usage of multiple scalabilities and perceptual quality estimators. We use H.264 scalable video coding in combined scalability mode to deliver videos to these devices. Using combined scalability increases the number of scalability options, so we need to perform multi-criteria optimization. In the low bitrate region, we show that multi-criteria optimization is not necessary and we propose an algorithm to reduce the complexity of the optimization problem. The rest of the chapter is as follows. In Section 4.1, we describe the system model. In Section 4.2, we compute the number of QEs that is required to perform optimum multicasting in our model. In Section 4.3, we describe the proposed approach which notably reduces the number of QE computations at low bitrates.

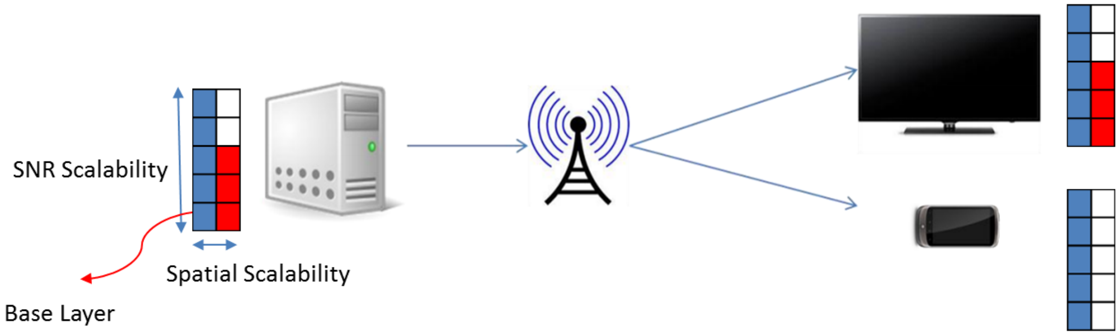


Figure 4.1: System Model

4.1 System Model

We represent the system model in Figure 4.1. In this figure, there is a content provider, a wireless access point and end users with different spatial resolution. We use H.264 Scalable Video Coding (SVC) [79] to compress the video into several layers. The first coded layer is the base layer and the remaining layers are enhancement layers. The decoding starts with the base layer and enhancement layers increase the quality gradually. SVC can be performed using three scalability options. They are SNR(Quality), Spatial and Temporal based scalabilities. We can also use more than one scalability option which is called Combined Scalability (CS) [80]. In this model, we use the combination of SNR (QP) and Spatial scalabilities.

We represent the SNR and Spatial layers of the encoded bitstream using a two-dimensional array as in Figure 4.1. The base layer is at the bottom-left first row and column of the array. All the other layers are enhancement layers. Each row denotes a different SNR scalability layer and each column shows a different spatial scalability layer. We can name a particular layer as $q_x q_y$ where x and y denote the number of SNR layers in the spatial base and spatial enhancement layers, respectively. We denote the total available transmission rate with R. The scalable bitstream consists of two parts. One part is decoded by both users. This part

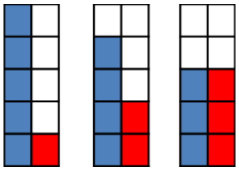


Figure 4.2: Combination of SNR layers at different spatial layers

is called the common part in our model and it is the left most column of the array. In Figure 4.1, the corresponding part of the bitstream is denoted with blue. The necessary rate to transmit this data is R_l. The second part of the bitstream, denoted by red blocks, shows the enhancement layers of spatial scalability. These blocks are only used by the receiver with the high resolution display. The necessary rate of this part is R -R_l.

During optimization, the aim is to allocate the available bitrate R to these blocks such that the quality is optimized for a specific metric. We denote the quality of the low-resolution user with Q_l and the quality of the high-resolution user with Q_h. Then, Q_l only depends on R_l. However, Q_h depends on both R_l and R because we can have different number of SNR layers in each spatial layer as shown in Figure 4.2. We can use the following two approaches to perform optimization as described in [34] and [35]:

1) Max-min Fairness : Maximize the minimum quality among users:

$$maxmin_{R,R_l}(N_h * Q_h(R, R_l), N_l * Q_l(R_l)) \tag{4.1}$$

2) Maximize the total utility of the users:

$$\max_{R,R_l}(N_h * Q_h(R, R_l) + N_l * Q_l(R_l)) \tag{4.2}$$

where N_h and N_l are the number of high resolution and low resolution video viewers, respectively. Perceptually optimum multicasting using combined scalability is a straight-forward problem, but it has two deficiencies. First, there is a dimension increase due to usage of an extra scalability option. The second one is the complexity of the QE used while computing the utility. In the next section, we describe how we compute the required number of QEs to perform optimum multicasting.

4.2 Complexity of Performing Optimum Multicasting Using Combined Scalability

The problem of multicasting to devices with different spatial resolutions is a multi-criteria optimization problem in [34, 35], so it needs to be solved based on a maximization or a fairness constraint. In either case, there is a trade off in quality between the devices of different spatial resolutions. We use combined scalability (CS) mode of H.264 SVC to compress the video such that we allocate the available bitstream to devices with different spatial resolutions flexibly.

In CS mode, we can have different SNR levels at different spatial layers. Hence, we need to compute the quality of each (SNR, Spatial) layer combination to perform the optimization. We assume that we have N layers in total and that we have 2 spatial layers. We also assume that the maximum number of layers for a upsampled low resolution (US-LR) video is N. For instance, in Figure 4.3, we show possible layer combinations when N is smaller than 7. Obviously, the total number of possible cases depends on N. The total number QE computations is calculated as follows:

If N is odd and $P = (N + 1)/2$,

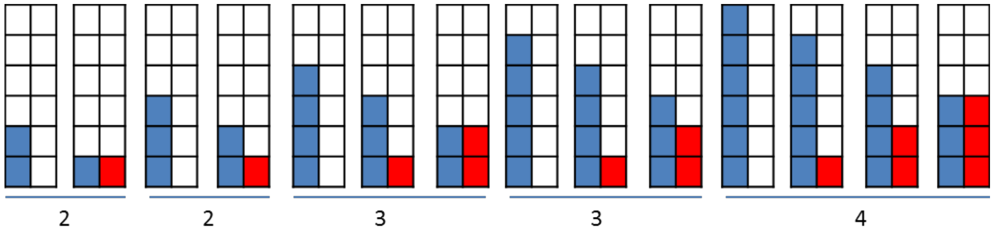


Figure 4.3: Possible layer combinations when the total number of layers is smaller than 7

$$TQ_o = 2 * (\frac{P * (P+1)}{2}) - 2 = \frac{N^2 + 4N - 5}{4} \tag{4.3}$$

On the other hand, if N is even and $P = N/2$, the total number of computations is:

$$TQ_e = \frac{P * (P+1)}{2} - 1 + \frac{P * (P+3)}{2} - 1 = \frac{N^2 + 4N - 4}{4} \tag{4.4}$$

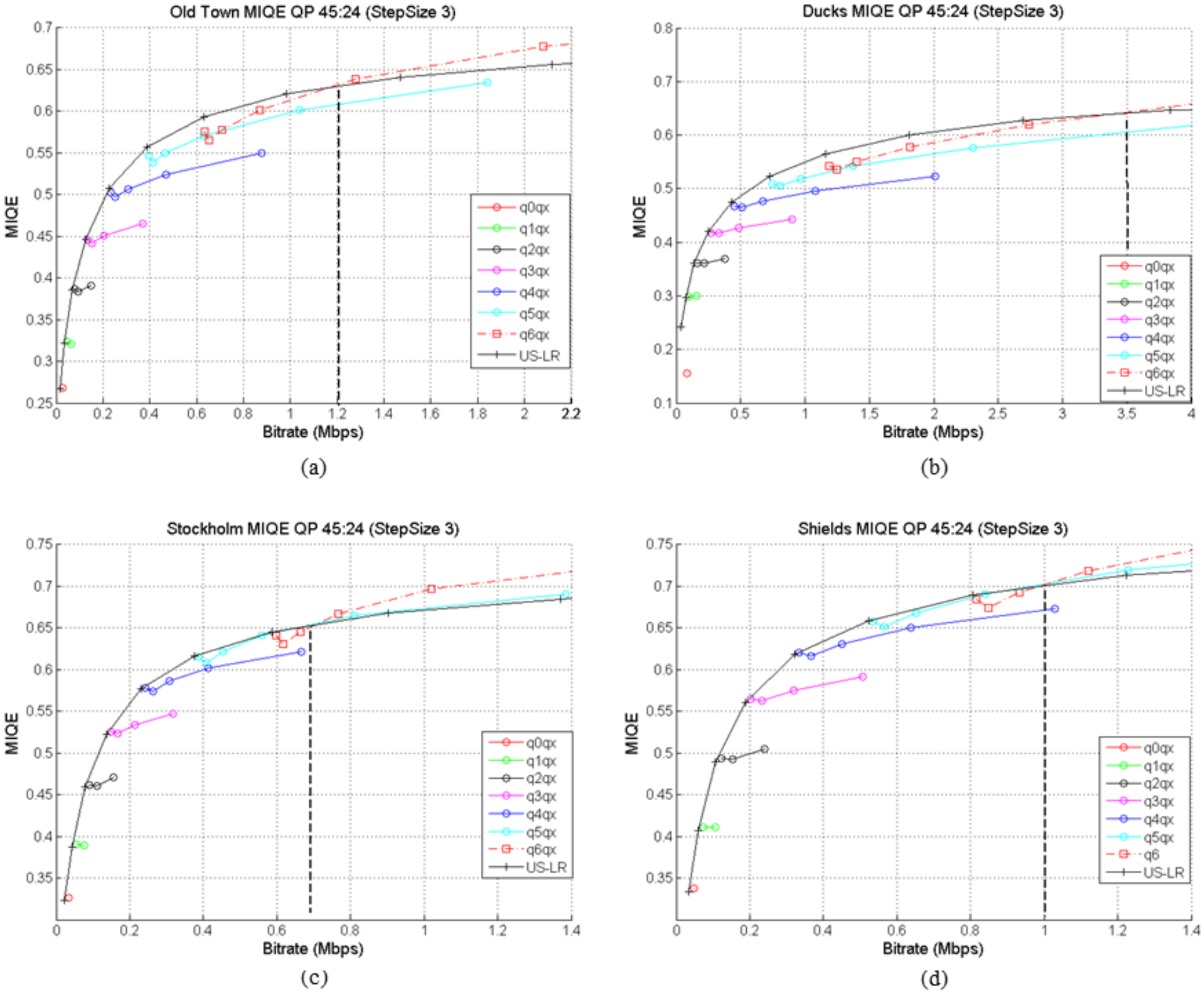
Hence, the total number of QE calculation is $O(N^2)$. Such a quadratic complexity may not be possible in real time. Furthermore, using such an optimization procedure is very time-consuming if we use a complicated perceptual QE approach like MOVIE [21].

4.3 Proposed Approach

In this section, we discuss the observations for optimum multicasting using combined scalabilities and how we use these observations to reduce the number of QE computations. We create scalable video streams using CS and compute their qualities to obtain. We have used H.264 SVC to encode videos in the combined scalability mode. We have used Old Town, Stockholm, Shields, and Ducks [67] videos. These videos have different spatial and temporal characteristics. The spatial resolution of the videos is 1280 × 720 and the frame rate of the videos is 50 fps. We use the normative option of JSVMs DownConvert application while decimating to create a low-resolution 640 × 360 video. To have reliable quality measurements, we should use a quality estimator which has a good performance for different spatial resolution videos that are displayed at the full resolution. Tables 3.7 and 3.9 in Section 3.3 and show the performance of QEs and their complexities with respect to PSNR for this task. According to these tables, MIQE and MVQE almost have the same performance but the complexity of MIQE is less than the complexity of MVQE. Hence, we use MIQE to estimate the quality. We compute the qualities of bitstreams for each CS combination. We also compute the qualities of US-LR videos which do not have any SNR layers in the spatial enhancement layer.

Figure 4.4 shows the rate-quality curves of CS and US-LR videos at low bitrates. In Figure 4.4, the colored curves with square symbols illustrate the rate-quality curves of CS videos. Each color denotes the number of SNR layers in the base spatial layer. For example, the blue curve for q4qx, depicts five circles representing the performance of a system with 5 layers in the spatial base and $x = 0, 1, 2, 3, 4$ spatial enhancement layers. The black curves with + symbols show the rate-quality curve of the US-LR videos. We have observed that the quality of the US-LR video is always better than any CS video at the same bitrate at low bitrates. This feature is valid up to a bitrate which we call the Critical Bitrate (CBR). The CBR is denoted with dashed lines in Figure 4.4. We can use this observation to reduce the complexity. At low bitrates, we downsample the video and compress it instead of only compressing it. In this case, we use all the available bitrate for the smaller resolution to maximize the quality of high-resolution video. This rate allocation scheme automatically maximizes the quality of the low-resolution video. Hence, it is not necessary to perform multi-criteria optimization at low bitrates. However, we can only use this property up to the CBR, so it is necessary to find CBR.

We can find the CBR by determining the intersection points between the quality-rate curve



(a)

(b)

(c)

(d)

Figure 4.4: Rate-Quality curves of CS and US-LR videos at low bitrates

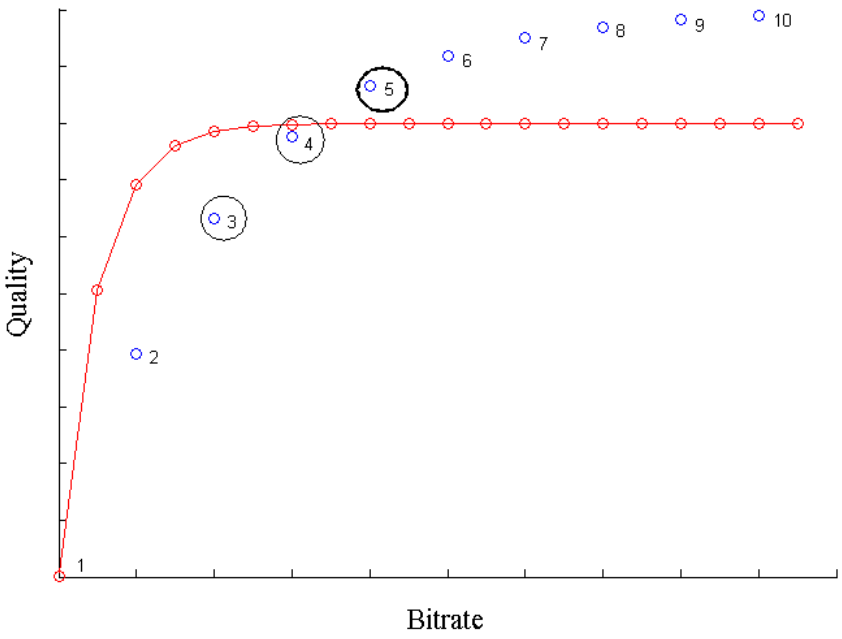


Figure 4.5: Binary search to find the SNR layer index of the spatial base layer

of the US-LR videos and the quality-rate curves of CS videos. The intersection point with the smallest bitrate is the CBR. Using this approach, the problem has evolved from finding the optimum allocation of layers for each bitrate to the computation of the CBR. We use Figures 4.5-4.6 to describe how we find the CBR. This approach consists of two parts. In the first part, we should determine SNR layer of the base spatial layer. Then, we need to determine the bitrate of the intersection point. Each colored curve represents a specific SNR layer in the base spatial layer in Figure 4.4. Our aim is finding the curve which intersects the rate-quality curve of the US-LR videos with minimum bitrate. We use a binary search to find the corresponding curve so we need $O(log_2(N))$ number of QE computations both for CS and US-LR videos as shown in Figure 4.5. In this figure, the blue circles denote the rate-quality values of CS videos which have the same number of SNR layers in both spatial base and enhancement layers. Moreover, the red curve denotes the rate-quality curve of the US-LR videos. In this part, we find the number of SNR layers in the base spatial layer and we denote this number with k. In Figure 4.5, k is equal to 5. Next we need to determine the

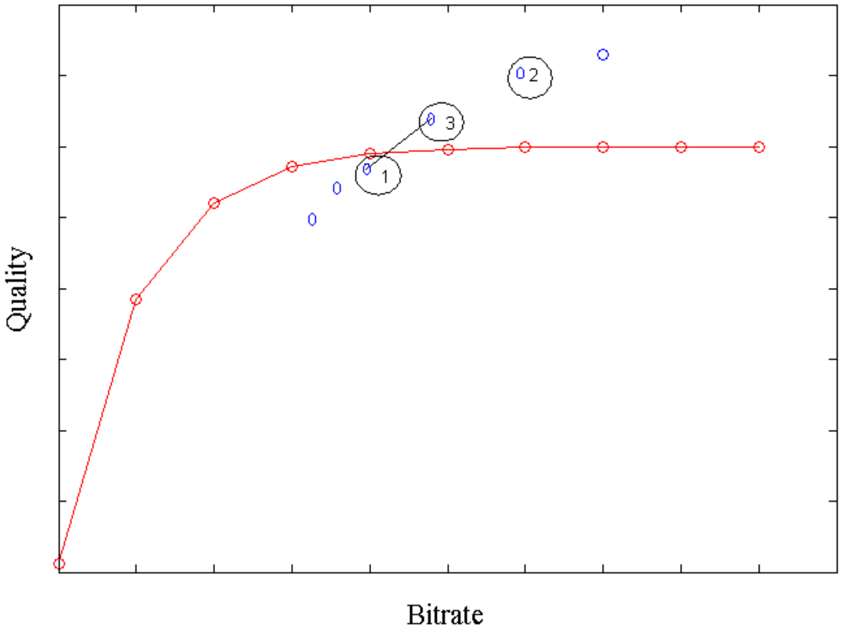


Figure 4.6: Binary search to find the SNR layer index of the spatial enhancement layer and the intersection point

critical bitrate. We search the SNR layer index in the spatial enhancement layer by using a binary search. In this case, the number of QE computations is $O(log_2(k))$ as shown in Figure 4.6. Hence the total number of computations is:

$$LTQ = 2 * (\lceil log_2(N) \rceil + \lceil log_2(k/2) \rceil) \leq 4 * (\lceil log_2(N) \rceil - 1) \tag{4.5}$$

As a result, the complexity drops from $O(N^2)$ to $O(log_2 N)$. This approach is especially useful when the stepsize of the QP is 1. For instance, in Figure 4.3b (Ducks), the intersection occurs when QP = 24 and the QP value of the base layer is 45. If the QP stepsize is 1, N will be 22. The total number of QE calculations using Eq. (4.4) is 139, and the total number of computations using the proposed approach is 20. Hence using this approach we can reduce the complexity of perceptually optimum multicasting at low bitrates without compromising the reliability of the perceptual QE.

Chapter 5

Conclusion

In this thesis we study the quality estimation and the delivery of videos with different spatial resolutions. First, we explore the quality estimation of images with resolutions different from that of the reference image. We show that conventional approaches have significant limitations to estimate this quality. We propose several ideas to overcome these limitations. We develop an algorithm using these ideas to solve the problem. We perform subjective tests to verify whether the proposed approach improves the results. We analyze the subjective test results using different test cases. When the reference and test images have different resolutions, the subjective tests demonstrate that in most cases the proposed method works better than other approaches. In addition, the proposed algorithm performs well when the reference image and the test image have the same resolution.

Second, we work on computing the quality of a video with lower spatial resolution compared to the reference video. We propose an algorithm to solve this problem. We develop this algorithm by incorporating the effect of motion into our previously proposed method for images, i.e. MIQE. We also perform subjective tests to measure the performance of the proposed approach. According to the test results, the proposed algorithm outperforms other

QEs when the reference video and test videos have different spatial resolutions. Lastly, we evaluate the performance of the proposed quality estimators using the Scalable Video Database.

Third, we work on the challenge of using a perceptual quality estimator to perform optimum multicasting of videos to the devices with different spatial resolutions. We specifically focus on the complexity of the optimum perceptual multicasting. The complexity increases due to usage of scalable video coding in combined scalability mode and perceptual quality estimators. Using combined scalability increases the number of scalability options, so we need to perform multi-criteria optimization. As a result of the simulations we have performed, we have observed that multi-criteria optimization is not necessary in the low bitrate region, and we propose an algorithm to reduce the complexity of the optimization notably for this region.

Bibliography

[1] (2013, Jun.) Cisco visual networking index: Forecast and methodology, 20122017. CISCO Corp. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html

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

[3] S. Chang and A. Vetro, "Video adaptation: concepts, technologies, and open issues," in *IEEE*, vol. 93, no. 1, January 2005, pp. 148–158.

[4] S. H. Bae, T. N. Pappas, and B.-H. Juang, "Subjective evaluation of spatial resolution and quantization noise tradeoffs," *IEEE Trans. Image Proc.*, vol. 18, pp. 495–508, March 2009.

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

[6] J. S. Lee, F. D. Simone, and T. Ebrahimi, "Subjective quality evaluation via paired comparison: Application to scalable video coding," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 882–893, May 2011.

[7] G. Cermak, M. Pinson, and S. Wolf, "The relationship among video quality, screen resolution, and bit rate," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 258–262, March 2011.

[8] J. McCarthy, M. Sasse, and D. Miras, "Sharp or smooth? comparing the effects of quantization vs. frame rate for streamed video," in *ACM CHI on Human Factors in Computing Systems*, April 2004, pp. 535–542.

[9] N. Cranley, P. Perry, and L. Murphy, "Optimum adaptation trajectories for streamed multimedia," *Multimedia Systems*, vol. 10, no. 5, pp. 392–401, August 2005.

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

[11] S. S. Hemami and A. R. Reibman, “No-reference image and video quality estimation: Applications and human-motivated design,” *Signal Processing: Image Communication*, August 2010.

[12] M. Saad and A. Bovik, “Blind prediction of natural video quality,” *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1352–1365, March 2014.

[13] A. Mittal, M. Saad, and A. Bovik, “Assessment of video naturalness using time-frequency statistics,” in *IEEE Int. Conf. Image Proc.*, October 2014, pp. 571 – 574.

[14] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, “A no-reference perceptual blur metric,” in *IEEE Int. Conf. Image Proc.*, September 2002, pp. 57–60.

[15] H. R. Wu and M. Yuen, “A generalized block impairment metric for video coding,” *IEEE Signal Processing Letters*, vol. 4, pp. 317–320, November 1997.

[16] A. K. Moorthy and A. C. Bovik, “A two-step framework for constructing blind image quality indices,” *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[17] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Proc.*, vol. 13, no. 4, pp. 600–612, April 2004.

[19] Z. Wang, E. Simoncelli, and A. Bovik, “Multiscale structural similarity for image quality assessment,” in *37th IEEE Asilomar Conference on Signals, Systems and Computers*, vol. 2, November 2003, pp. 1398–1402.

[20] M. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality,” *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, September 2004.

[21] K. Seshadrinathan and A. C. Bovik, “Motion tuned spatio-temporal quality assessment of natural videos,” *IEEE Transactions on Broadcasting*, vol. 19, no. 2, pp. 335–350, February 2010.

[22] S. Daly, “Engineering observations from spatiovelocity and spatiotemporal visual models,” in *Human Vision and Electronic Imaging III*, July 1998, pp. 180–191.

[23] R. Feghali, D. Wang, F. Speranza, and A. Vincent, “Quality metric for video sequences with temporal scalability,” in *IEEE Int. Conf. Image Proc.*, September 2005, pp. 137–140.

[24] Y. Peng and E. Steinbach, “A novel full-reference video quality metric and its application to wireless video transmission,” in *IEEE Int. Conf. Image Proc.*, September 2011, pp. 2517–2520.

[25] Y. Xue, Y.-F. Ou, Z. Ma, and Y. Wang, "Perceptual video quality assessment on a mobile platform considering both spatial resolution and quantization artifacts," in *Packet Video Workshop*, December 2010.

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

[27] C. S. Kim, S. H. Jin, D. J. Seo, and Y. M. Ro, "Measuring video quality on full scalability of H.264/AVC scalable video coding," *IEICE Transactions on Communications*, no. 5, pp. 1269–1278, October 2010.

[28] H. Sohn, H. Yoo, W. D. Neve, C. S. Kim, and Y. M. Ro, "Full-reference video quality metric for fully scalable and mobile SVC content," *IEEE Transactions on Broadcasting*, vol. 56, no. 3, pp. 269–280, September 2010.

[29] J. S. Lee, "Comparison of objective quality metrics on the scalable extension of H.264/AVC," in *IEEE Int. Conf. Image Proc.*, October 2012, pp. 693–696.

[30] A. Besson, F. De Simone, and T. Ebrahimi, "Objective quality metrics for video scalability," in *IEEE Int. Conf. Image Proc.*, 2013, pp. 59–63.

[31] J. Vella and S. Zammit, "A survey of multicasting over wireless access networks," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 2, pp. 718–753, June 2012.

[32] I. Lee, J. Park, S. Kim, T. Oh, and S. Lee, "Device-aware visual quality adaptation for wireless n-screen multicast systems," *IEICE Transactions on Communications*, no. 12, pp. 3181–3189, December 2013.

[33] W. Ji, Z. Li, and Y. Chen, "Joint source-channel coding and optimization for layered video broadcasting to heterogeneous devices," *IEEE Transactions on Multimedia*, vol. 14, no. 2, pp. 443–455, April 2012.

[34] S. P. Chuah, Z. Chen, and Y. P. Tan, "Energy-efficient resource allocation and scheduling for multicast of scalable video over wireless networks," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1324 – 1336, August 2012.

[35] S. Luo, S. Sun, X. Jing, Y. Lu, and N. Chen, "On using cooperative game theory to solve the wireless scalable video multicasting problem," in *ISCC*, July 2013, pp. 318 – 323.

[36] C. Chan, S. Wee, and J. Apostolopoulos, "Multiple distortion measures for packetized scalable media," *IEEE Transactions on Multimedia*, vol. 10, no. 8, pp. 1671 –1686, Dec. 2008.

[37] E. Reed and J. Lim, "Optimal multidimensional bit-rate control for video communication," *IEEE Trans. Image Proc.*, vol. 11, no. 8, pp. 873 – 885, Aug. 2002.

[38] E. Akyol, A. M. Tekalp, and M. R. Civanlar, “Content-aware scalability-type selection for rate adaptation of scalable video,” *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, 2007.

[39] A. M. Demirtas, H. Jafarkhani, and A. R. Reibman, “Quality estimation for images and video with different spatial resolutions,” in *Human Vision and Electronic Imaging XVII*, vol. 8291, Feb. 2012.

[40] K.N.Ngan, K. Leong, and H. Singh, “Adaptive cosine transform coding of images in perceptual domain,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1743–1750, November 1989.

[41] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” *IEEE Trans. Image Proc.*, vol. 15, no. 2, pp. 430–444, February 2006.

[42] E. C. Larson and D. M. Chandler, “Most apparent distortion: Full-reference image quality assessment and the role of strategy,” *Journal of Electronic Imaging*, vol. 19, no. 1, March 2010, http://vision.okstate.edu/index.php?loc=csiq.

[43] D. H. Kelly, “Motion and vision II: Stabilized spatio-temporal threshold surface,” *J. Opt. Soc. Am.*, vol. 69, no. 10, pp. 1340–1349, October 1979.

[44] D. M. Chandler and S. S. Hemami, “VSNR: A wavelet-based visual signal-to-noise ratio for natural images,” *IEEE Trans. Image Proc.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.

[45] A. B. Watson, “The cortex transform: Rapid computation of simulated neural images,” *Comput. Vision Graph. Image Process.*, vol. 39, no. 3, pp. 311–327, Sep. 1987.

[46] S. Daly, “The visible differences predictor: an algorithm for the assessment of image fidelity,” in *Digital images and human vision*, A. B. Watson, Ed. Cambridge, MA, USA: MIT Press, 1993, pp. 179–206.

[47] A. Cohen, I. Daubechies, and J. C. Feauveau, “Biorthogonal bases of compactly supported wavelets,” *Communications on Pure and Applied Mathematics*, vol. 45, no. 5, pp. 485–560, June 1992.

[48] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, “On advances in statistical modeling of natural images,” *Journal of Mathematical Imaging and Vision*, vol. 18, pp. 17–33, 2003.

[49] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, “Random cascades on wavelet trees and their use in analyzing and modeling natural images,” *Applied and Computational Harmonic Analysis*, vol. 11, pp. 89–123, 2001.

[50] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* Wiley-Interscience, 2006.

[51] D. H. Kelly, “Spatiotemporal variation of chromatic and achromatic contrast thresholds,” *J. Opt. Soc. Am.*, vol. 73, no. 6, pp. 742–750, 1983.

[52] D. M. Rouse, R. Pépion, S. S. Hemami, and P. L. Callet, “Image utility assessment and a relationship with image quality assessment,” in *Human Vision and Electronic Imaging*, 2009.

[53] C. A. Segall and G. J. Sullivan, “Spatial scalability within the H.264/AVC scalable video coding extension,” *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 17, no. 9, pp. 1121–1135, Sep 2007.

[54] A. R. Reibman, “A strategy to jointly test image quality estimators subjectively,” in *Proc. International Conference on Image Processing*, 30 Sept–3 Oct. 2012.

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

[56] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, “Complex wavelet structural similarity: a new image quality index,” *IEEE Trans. Image Proc.*, vol. 18, no. 11, pp. 2385–2401, 2009.

[57] L. L. Thurstone, “A law of comparative judgment,” *Psychological Review*, vol. 34, pp. 273–286, 1927.

[58] R. A. Bradley and M. E. Terry, “The rank analysis of incomplete block designs — I. The method of paired comparisons,” *Biometrika*, vol. 39, pp. 324–345, 1952.

[59] J. C. Handley, “Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment,” in *Image Proc, Image Qual., and Image Capture Sys. Conf. (PICS’01)*, 2001, pp. 108–112.

[60] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, “A statistical evaluation of recent full reference image quality assessment algorithms,” *IEEE Trans. Image Proc.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[61] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, “TID2008 - A database for evaluation of full-reference visual quality assessment metrics,” *Advances of Modern Radioelectronics*, vol. 10, pp. 30–45, 2009.

[62] A. M. Demirtas, A. R. Reibman, and H. Jafarkhani, “Full-reference quality estimation for images with different spatial resolutions,” *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2069–2080, May 2014.

[63] A. B. Watson, J. Hu, and J. F. M. III, “DVQ: A digital video quality metric based on human vision,” *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 335–350, March 2001.

[64] J. You, T. Ebrahimi, and A. Perkis, “Attention driven foveated video quality assessment,” *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 200–213, January 2014.

[65] A. B. Watson and A. J. Ahumada, “Model of human visual-motion sensing,” *Journal of The Optical Society of America*, vol. 2, no. 2, pp. 322–341, February 1985.

[66] Y.-Q. Zhang and S. Zafar, "Motion-compensated wavelet transform coding for color video compression," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 2, no. 3, pp. 285–296, September 1992.

[67] Xiph.org video test media. http://media.xiph.org/video/derf/.

[68] Rec ITU-T P.910, "Subjective video quality assessment methods for multimedia applications," International Telecommunication Union, Tech. Rep., 2006.

[69] JVT. Reference software version JM18.1. http://iphome.hhi.de/suehring/tml/download/oldjm/jm18.1.zip.

[70] S. Sun and J. Reichel, "AHG report on spatial scalability resampling," in *ISO/IEC JTC1/SC29/WG11, Doc JVT-R006*, 14 – 20 Jan. 2006.

[71] "Multiscale Video Quality Estimation," 2015. [Online]. Available: http://cpcc.uci.edu/mvqe.php

[72] S. Péchard, R. Pépion, and P. Le Callet, "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm," in *International Workshop on Image Media Quality and its Applications, IMQA2008*, Kyoto, Japan, September 2008, pp. 6–12.

[73] Rec. ITU BT.1788, "Methodology for the subjective assessment of video quality in multimedia applications," International Telecommunication Union, Tech. Rep., 2007.

[74] Rec. ITU-R BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Tech. Rep., 2002.

[75] G. W. Snedecor and W. G. Cochran, *Statistical Methods.* Iowa State University Pres, 1989.

[76] Rec. ITU P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," International Telecommunication Union, Tech. Rep., 2012.

[77] D. A. Kenny, *Statistics for the Social and Behavioral Sciences.* Little, Brown, 1987.

[78] V. Q. E. G. (VQEG), "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment," Mar. 2000. [Online]. Available: ftp://vqeg.its.bldrdoc.gov/Documents/Projects/frtv1/other/COM-80E_final_report.pdf

[79] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Ckts. Syst. for Video Tech.*, vol. 17, no. 9, pp. 1103 – 1120, September 2007.

[80] H. Schwarz, D. Marpe, T. Schierl, and T. Wiegand, "Combined scalability support for the scalable extension of H.264/AVC," in *IEEE Int. Conf. on Multimedia and Expo*, July 2005, pp. 446 – 449.