PERCEPTUAL VIDEO QUALITY ASSESSMENT AND ANALYSIS USING ADAPTIVE CONTENT DYNAMICS

A Dissertation
Presented to
The Academic Faculty

by

Mohammed A. Aabed

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
May 2017

PERCEPTUAL VIDEO QUALITY ASSESSMENT AND ANALYSIS USING ADAPTIVE CONTENT DYNAMICS

Approved by:

Professor Ghassan I. AlRegib, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Elliot Moore II
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Anthony J. Yezzi
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Berdinus A. Bras
The George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Professor David V. Anderson
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: December 21st, 2016

In the memory of my mother,

to my father,

and

to my brothers

ACKNOWLEDGEMENTS

I would like to express my utmost gratitude to my advisor, Prof. Ghassan AlRegib, a man whose guidance, patience and generosity have always made him a trusted mentor and a dear close friend. His continuous teaching, friendship, support, and understanding throughout this process have allowed me to learn, innovate, enjoy and made it a rewarding experience. I would like to thank Dr. Anthony J. Yezzi and Prof. David V. Anderson for their time, valuable comments and for agreeing to serve as reading remember of my dissertation committee. I would like to also thank Prof. Elliot Moore II and Prof. Bert Bras for their time, comments, and for agreeing to serve on my defense committee.

This dissertation would not have seen the light without a great working environment of the Multimedia and Sensors Lab (MSL) and Center for Signal and Information Processing (CSIP). I thank all CSIP faculty, staff, and students. I also would like to thank all my friends and colleagues within MSL and CSIP. Special thanks go to my dear friends and colleagues, Dr. Dogancan Temel, Dr. Mashhour Solh, Dr. Sami Almalfouh and Ms. Patricia Dixon for their valuable friendships, collaborations, and support over the years. My thanks go to the academic office staff at the School of Electrical and Computer Engineering at the Georgia Institute of Technology.

I am forever in debt to my family for their endless care, encouragement, patience, and unconditional support. My everlasting gratitude goes to my first and foremost teacher, my mother, for her teachings, guidance, care and dedication. I am eternally grateful to my father for the continuous dedication, care, support and wise guidance. Special thanks go to my brother, Dr. Mohanad Aabed, MD, for his unconditional support and encouragement over the years. My gratitude goes also to my brothers,

Mr. Khaled Abed and Mr. Saed Aabed, for their support and encouragement.

I would like to thank all my friends who supported me through this endeavor. Special thanks go to Dr. Osama Mohamad, MD, for his valuable friendship and support over the years.

"In the sweetness of friendship let there be laughter, and sharing of pleasures. For in the dew of little things the heart finds its morning and is refreshed."[1] To all of you my family, teachers, friends, and colleagues, I humbly thank you.

[1]Khalil Gibran

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

With the growth of mobile data services and bandwidth, several applications and streaming services have emerged that made video quality and technologies important fields of research and development. Understanding perceptual video quality can be achieved through understanding and tightly linking the perceptual nature of the human visual system and varying characteristics and dynamics of video contents.

In this dissertation, the objective of the proposed research is to investigate perceptual quality assessment and analysis of videos subject to different types of distortion. We propose utilizing adaptive content dynamics to examine the impact of different error sources on the perceptual quality of the video. We design perceptual video quality estimators using novel handcrafted features inspired by the human visual properties. We explore new feature spaces and utilize them to capture varying video dynamics as experienced by our visual perception. Specifically, we introduce a new framework for perceptual video quality using pixel-level optical flow maps where we propose a motion processing procedure inspired by the hierarchical processing of motion in the visual cortex. Furthermore, we propose another perceptual video quality assessment approach by examining the varying properties of the tempospatial power spectrum. Using the power spectrum, we design a novel sensitivity measure to capture the impact of distortions on visual perception. This work includes a full-reference computationally efficient framework that captures both spatial and temporal characteristics in the frequency domain. We also examine the performance of various statistical moments and pooling strategies, at both spatial and temporal levels, with different visual feature maps. This aims at revealing the optimal pooling strategies most correlated with visual perception for every feature space with respect to different distortions.

CHAPTER I

INTRODUCTION

Over the past decade, video streaming services have gained a massive popularity and both content and the number of users are continuously growing. The continuous growth of Internet traffic in general, and video traffic in particular, has triggered the signal processing and communication communities concern with bandwidth and quality of experience (QoE). Global IP traffic has increased significantly over the past years and is expected to continue growing over the next few years. Mobile data traffic has increased 4000-fold over the past decade. It is also predicted to reach 30.6 exabytes per month by 2020 (over eightfold increase from 2015), out of which 75% will be video traffic (11-fold increase from 55% in 2015) [16]. Average mobile connection speed is expected to grow from 2 Mbps in 2015 to 6.2 Mbps in 2020 [17]. Mobile video represented more than half of global mobile data traffic beginning in 2012, indicating that it is already affecting traffic today, not just in the future. A million minutes of video content is estimated to cross the network every second by 2018 [17]. It is estimated that a growth of 68% in global mobile connections will occur by 2020 reaching 11.6 billion mobile connections. Mobile video traffic will account for over 75% of that total. Furthermore, busy-hour[1] Internet traffic is growing more rapidly than average Internet traffic [16, 17]. Thus, the standardization bodies are adapting to this growth by motivating technologies that increase the efficiency of bandwidth utilization, data compression and QoE.

Nevertheless, as mobile video traffic increases rapidly and streaming technology adapts to comply with the demand, the importance of quality of experience (QoE)

[1]Busy-hour is the busiest 60-minute period in a day [16].

has been more emphasized. A survey published in 2015 revealed that one out of five viewers will abandon a poor streaming service immediately while 75% will tolerate a bad stream for up to four minutes [18]. To establish stable video streaming networks while maintaining high quality of the videos, perceptual video quality assessment (PVQA) becomes an essential research topic in the video communication society. To motivate this topic and research, we introduce an example to show the impact of coding operations on data dependency in recent video coding standards on video quality. Furthermore, we show the impact of network errors or losses under such coding conditions. As it will be explained in details in Chapter 2, H.246/MPEG-4 Advance video coding (AVC) is the most common video coding standard in active systems. Its successor, H.265/MPEG-H High Efficiency Video Coding (HEVC) standard, was introduced in 2013. The design of HEVC standard included many new features to efficiently enable random access and bitstream splicing. Many functionalities such as channel switching, seeking operations, and dynamic streaming services require a good support of random access. These features, however, make the bitstream and the decoded sequence more sensitive to errors and losses due to the higher level of data dependency. This, in turn, introduces more challenges in terms of video quality assessment and monitoring, error concealment, etc. To this end, we examine the impact of channel errors or losses on the fidelity of the decoded HEVC video by estimating the channel-induced distortion.

HEVC employs an open Group of Picture (GOP) format in which inter-coded pictures (temporal redundancy) are used more frequently than AVC to allow higher compression gain. In contrast to H.246/MPEG-4 AVC, H.265/MPEG-H HEVC employs an open GOP operation. In this format, a new clean random access (CRA) picture syntax is used wherein an intra-coded picture (spatial redundancy) is used at the location of random access point (RAP) to facilitate efficient temporal coding [12]. The intra period varies depending on the frame rate to introduce higher

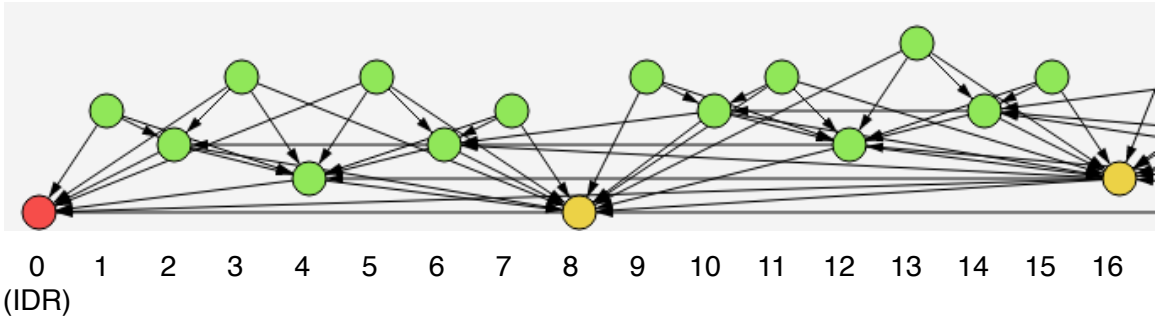


Figure 1: The open GOP structure in HEVC coded videos [1]

.

compression gain [19]. This coding structure is shown in Fig. 1. In this figure, frames are represented using circles and the order at the bottom of the figure is the picture order count (POC). The Red circle represents a intra-coded frame (I-frame), where the frame is compressed utilizing *spatial redundancy only* and independent of other frames. Yellow circles represent inter-coded frames (P-frames), where frames are compressed exploiting both spatial and temporal redundancies in I- and P-frames. Green circles represent another type of inter-coded frames (B-frames), where frames are compressed exploiting both spatial and temporal redundancies in other I-, P-, and B-frames. The picture quality and packet size are the highest for I-frames, followed by P-frames and B-frames, respectively[2]. The sequence starts with an I-frame (POC 0) which is followed by a P-frame (POC 8) and 7 B-frames (POCs 2 through 7) to form an open GOP of size 8. The next open GOP starts with the P-frame (POC 8) from the previous GOP (frames 8-16 in Fig. 1). This pattern continues until the end of the intra period. The arrows in the figure represent decoding dependencies.

In HEVC, favouring inter-coding over intra-coding is more subtle than in AVC. As a result, HEVC imposes a very high data dependency between the frames. Henceforth, the impact of channel-induced errors on certain frames that potentially propagate to the end of the GOP is more significant in HEVC than in AVC. Fig. 2 shows an example

[2]The details and axioms of video coding and compression will be discussed in details in Sections 2.1.1.2 and 2.1.2

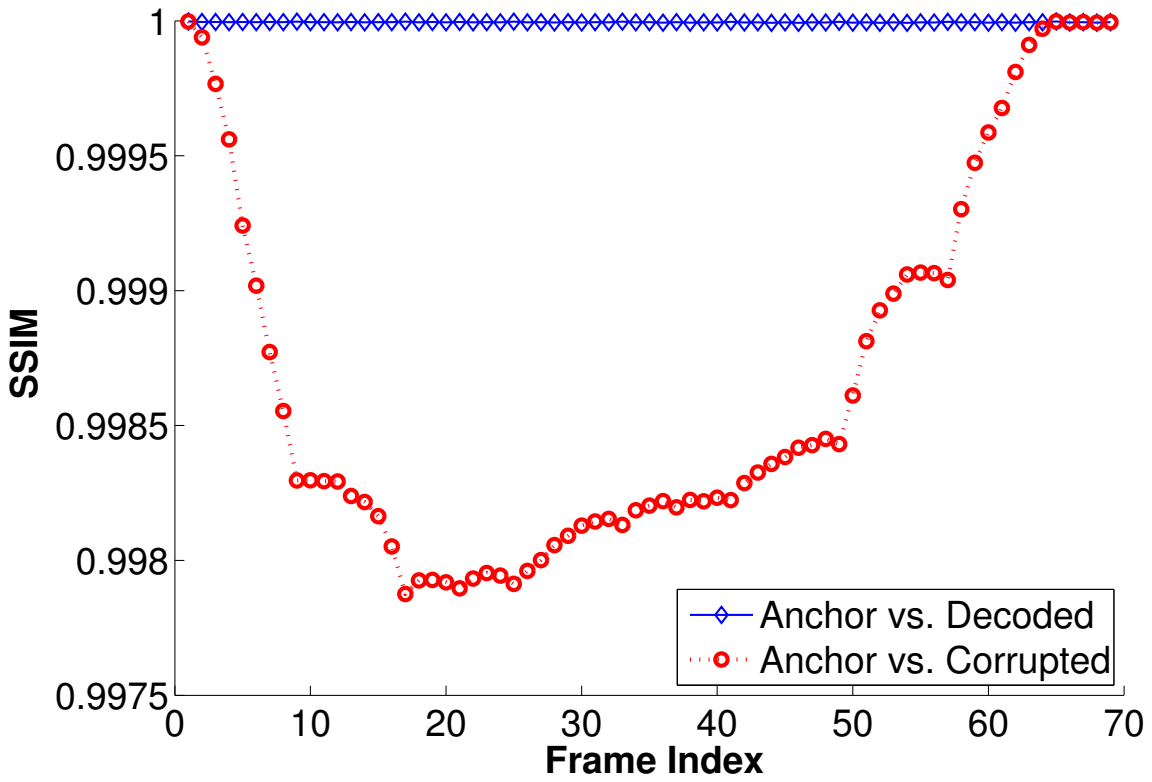


Figure 2: The impact of loosing frame 8 on the SSIM values of the GOP for `BQMall` sequence; frame rate is 60 frames per second.

of the impact of loosing the Network Abstraction Layer (NAL) unit corresponding to frame 8 and replacing it with the temporally closest available frame at the decoder, which is frame 0 in this example (See Fig. 1). In our simulations and tests, we abide by the recommended encoding format wherein every frame is taken as a single slice which is encapsulated in a separate NAL unit [20]. Fig. 2 shows that the channel loss under these coding conditions propagates until a new I-frame is encountered, which is frame 64 in this example.

In light of these concerns, the issue of primary focus in this dissertation is the impact of all these operations and sources of distortion on the end user's *perceived quality of experience*. Under the assumption that we do not have access to the decoder and we only have access to the decoded pictures, we do not have knowledge of how losses have propagated to other frames. Furthermore, for some error concealment

techniques, it might be hard to measure the propagated error through the traces of temporal error concealment, as proposed in [21, 22]. Hence, in order to estimate these distortions, we can only rely on the spatial and temporal features of the decoded video.

The goal in this work is to assess, understand, and analyze video quality, its characteristics, and its changing features under different coding, streaming and content conditions. The variation in contents and coding parameters cause a change in the dynamics that affect the perceived quality. We tackle this problem by studying the variations in content dynamics and features tempospatially. In streaming applications, coding parameters and bitstream format vary depending on the application and compression standard. Additionally, the content provider in some applications could encrypt the bitstream or deny access to it. Consequently, it is not possible to examine the bitstream and its metadata. In such cases, video quality can only be evaluated by pixel-based methods which evaluate the video contents directly. Furthermore, it is not possible to accurately examine the perceptual quality of visual media without examining perceptual stimuli in video contents, including spatial, temporal, frequency and any visually correlated characteristics. The examination of the impact of video contents and visual features on visual perception can be best achieved by analyzing pixel-level features. In this dissertation, we investigate different features and criteria for perceptual video quality assessment focusing on both spatial and temporal features. We propose algorithms and techniques towards this end exploring new video properties and novel features spaces.

1.1 Contributions and Dissertation Organization

The contributions of this dissertation can be summarized as follows:

1. In Chapter 3, we propose a perceptual video quality assessment approach using optical flow-based distortions maps. This is a reduced-reference perceptual

video quality metric to estimate distortion due to compression and network losses. The proposed technique does not make any assumption about the coding conditions or video sequence. It rather explores the temporal changes between the frames by analyzing the variations in the statistical properties of the optical flow.

2. In Chapter 4, we propose utilizing power spectral analysis to estimate the perceptual quality of videos. This chapter includes two new perceptual quality metrics. We start by designing a low-complexity no-reference video quality measure to estimate the channel-induced distortion at the frame-level due to network losses. Secondly, we propose a perceptual objective quality assessment framework based on tempospatially unified power-spectral density characteristics. This is a full-reference perceptual video quality assessment metric for distorted videos by analyzing the power spectral density. This estimation approach relies on the changes in video dynamic calculated in the frequency domain and are primarily caused by distortion.

3. In Chapter 5, three distortion maps are analyzed, spatially and temporally, to identify the most effective statistical moments and pooling strategies with respect to PVQA. The three distortion maps examine three visual feature: pixel fidelity, local structural similarity and motion fields. We show the most significant spatial and temporal features correlated with perception for every distortion map with respect to different distortion types. We use this data to draw insights about the human perception and its sensitivity to distortion. We also demonstrate that the same distortions across databases yield different results in terms of PVQA evaluation and verification. This work reveals the necessity for a verification and validation framework for PVQA databases.

The rest of this dissertation is organized as follows. Following the motivational introduction and problem description, Chapter 2 introduces the necessary background information about video coding, visual perception and perceptual video quality assessment. This part includes a comprehensive literature survey spanning all prior arts in this domain. Chapters 3, 4, and 5 introduce the novel contributions of this dissertation. Chapter 3 introduces a new framework for PVQA using pixel-level optical flow maps. Chapter 4 introduces the details of a proposed approach to video quality assessment by examining the varying dynamics of the tempospatial power spectrum. This chapter includes two novel algorithms: the first is a no-reference low-complexity metric for streaming applications, and the second is a general full-reference framework to perceptual quality assessment utilizing tempospatially unified power spectra. In Chapter 5, we examine the performance of various statistical moments and pooling strategies, at both spatial and temporal levels, with different visual feature maps. Finally, Chapter 6 details the conclusion remarks and future plans of this work.

CHAPTER II

BACKGROUND AND PRIOR ART

This chapter starts by introducing the fundamentals and necessary background related to video coding, compression and distortions. This is followed by an overview of the human visual system and the relevant characteristics to PVQA. In Section 2.4, the fundamentals and background of perceptual video quality assessment are discussed in details. This is followed by a thorough literature survey of prior arts in this domain. The survey concludes by highlighting the contributions in this dissertation and their novelties with respect to prior art.

2.1 Digital Video Streaming and Compression

Since the eighties of the past century, various standards and tools of video compression have been developed. The goal has been to maximize compression rate while maintaining video visual quality. Until recently, two organizations dominated video compression standardization. The first was the International Telecommunications Union Telecommunications Standardization Sector (ITU-T) Video Coding Experts Group (VCEG). The second was the International Standardization Organization and International Electrotechnical Commission (ISO/IEC) Moving Picture Experts Group (MPEG) [2]. The ITU-T developed a series of coding standards starting with H.261, H.262, and H.263 standard [23]. The MPEG also developed a number of standards including MPEG-1, MPEG-2, MPEG-3 and MPEG-4. Then, a joint collaboration forming the Joint Video Team (JVT) resulted in the H.264/MPEG-4 AVC standard [24]. MPEG and VCEG went on to establish the Joint Collaborative Team on Video Coding (JCT-VC), which developed the High Efficiency Video Coding (HEVC) standard [12, 25]. Before we discuss the history and advances in video coding, we

discuss the essentials and axioms of digital video compression and coding. This is followed by an overview of the video coding standards developments and history. To the best of the author's knowledge, this overview is the first of its kind in terms of comprehensiveness spanning all efforts from standardizations bodies, industry and major players worldwide. This aims to highlight the technological evolution and market competition in this rich domain which are major motivations behind this work [1].

2.1.1 Video Coding Fundamentals

2.1.1.1 Scope of Coding Standardization

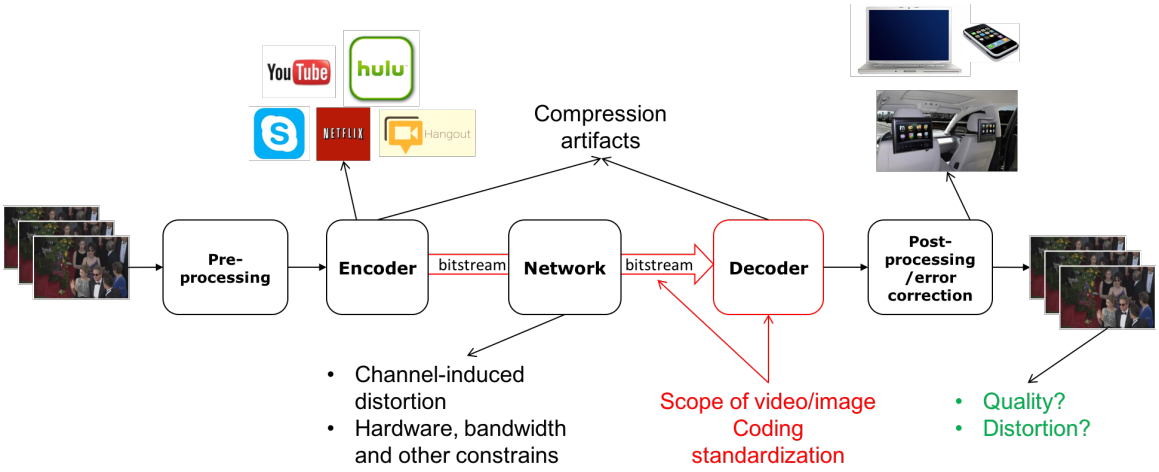


Figure 3: The basic components of video streaming system.

Figure 3 illustrates the basic components of a video streaming system. Given a sequence of time-moving pictures, the encoder performs a lossy compression operation to encode video pictures in a bitstream. During transmission, the bitstream is subject to different sources of errors and erasures inducing distortions to the decoded video. Upon receiving packets, a decoder performs a reconstruction of the time-moving pictures from the received, possibly corrupted, bitstream. The post-processing and error

[1] The survey introduced here focuses on video technology without discussing the details of audio development, which is out of the scope of this dissertation.

concealment modules then try to enhance the quality of the decoded video and compensate for different distortions. This process yields a best-effort video sequence.

Video standards targets establishing a universal framework and format for video streaming, storage and recording maintaining compatibility across devices, platforms, networks, etc. Thus, video standard development process goes through the following steps [26]:

- Identification of requirements
- Developments phase
- Selection of basic methods
- Collaboration phase
- Draft international standard
- Validation phase
- International standard

The scope of video and image coding standardization defines only *bitstream format (syntax) and decoder* of video codec. This framework does not specifically define the encoder design but defines the output format. A standard also defines several tools and components for compression, not all of which are required to produce a complaint bitstream. Only a subset of these tools must be implemented in the decoder to ensure universal comparability. This allows researchers, developers and architects room for optimization, complexity reduction for implementability and different applications [2, 26]. *This flexibility however comes with no guarantees of quality, perceptual or otherwise. In fact, this design makes predicting and modeling compression artifact and subsequent distortions very hard to predict and characterize.*

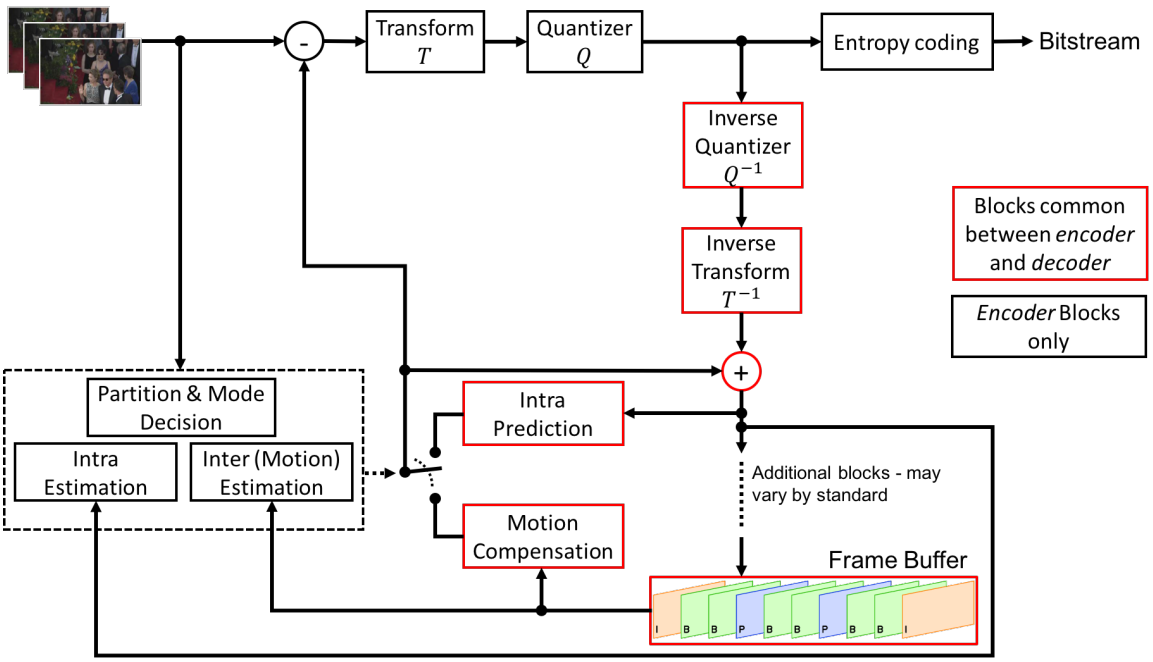


Figure 4: The basic components of a transform video coding system.

2.1.1.2 Video Coding Operations

The basic components a transform video compression are shown in Figure 4. This digram shows the basic blocks in the encoder and decoder. The fundamental operations of a transform video coding system are:

Partition and Mode Decision: When a video frame or slice is encoded, it may be compressed utilizing spatial redundancy, in which case the frame is known as an intra-frame (I-frame). An I-frame is compressed spatially and without any dependency on other frames. Alternatively, a frame maybe compressed exploiting temporal redundancy, in which case the frame is known to be an inter-frame (P-frame or B-frame). The compression ratio of inter-frames is significantly higher than that for intra-frames. This is because exploiting temporal redundancy allows for more flexibility in terms of search for redundant contents. The motion compensation module searches for similar contents in reference frames and only differential data (residuals) are compressed. Figure 4 shows the processing path

for each coding mode. Inter-frames have two compression modes. It may be compressed using forward prediction (P-frame) using only previous frames, I- or P-frames. Alternatively, a B-frame is compressed using bi-directional prediction from both earlier and later frames. As a result, B-frames have higher compression ratios than P-frames [23].

For any given frame to be encoded, the search range is confined to a set of candidate slices in the decoded frames buffer with which block-based correlation is maximal. Then, the closest matching blocks, also known as reference slices, are chosen for prediction and compensation.

Prediction and Compensation: After choosing the coding mode for currently encoded frame (slice), the encoder utilizes spatial and temporal redundancy in-between slices to reduce the amount of transmitted data. This is performed by calculating (*i*) the difference between encoded frame and the reference block, and (*ii*) the motion vector accounting for the tempospatial displacement of the predicted block. This process yields a residual frame and a set of motion vectors for each block in the predicted frame. Further operations maybe performed to reduce the energy of the residuals making this signal more sparse to enhance the compression efficiency This operation is known in video coding as *motion compensation*. As a result, the transmitted bitstream contains only the resultant residual frame and motion vectors for any encoded frame. At the receiver side, these operations are reversed by the decoder to reconstruct the predicted frame.

Transformation: After a frame is intra- or inter-compensated, the residuals are transformed to the frequency domain for further redundancy elimination and decorrelation. The human visual system has a higher sensitivity to certain frequency bands than others. This implies that the contents falling in frequency

ranges to which the HVS is more sensitive will have a higher impact on visual perceptual quality. This fact is utilized in video coding to allocate separate between contents in different frequency bands of different significance to the HVS. Then, more resources and bits are allocated to contents in more visually significant frequency bands. As a results, the compression efficiency is improved without significantly compromising visual quality. The discrete cosine transform (DCT) is the most common frequency domain transformation used for this purpose in video coding. DCT facilitates the basic three requirements a transform must satisfy for video coding operations which are reversibility, computational feasibility, and facilitation of data decorrelation [27]. The first two are of higher significance at the decoder where this transformation operation is reversed to reconstruct (decode) images. The third requirement enables the visually inspired processing alluded earlier.

Quantization: Following transformation, all video data is quantized to improve coding efficiency. The level of quantization varies depending on several aspects including available bandwidth and resources, application and content delivery requirements, visual quality requirements, etc. The level of quantization is controlled using a *quantization parameter (QP)*. Higher QP values correspond to more coarse quantization. In other words, the QP value is inversely proportional to the resultant bitrate and visual video quality.

Entropy Coding: This is a form of compression where the encoder takes advantage of the fact that certain symbols are more likely to exist in the transmitted data. The operations of entropy coding, also known as variable length coding (VLC), are generally independent of the medium characteristics. The encoder assigns variable-length codewords to each symbol [27]. Two popular approaches to entropy coding are Hoffman and arithmetic coding, where arithmetic coding

may be viewed as a generalized form of Hoffman coding. Arithmetic coding encodes the entire message into one number represented in a single arithmetic base where Hoffman codes represent every symbol of the message using a series of digits in the arithmetic base. As a result, general arithmetic coding sometimes reaches optimal entropy encoding much more closely than Huffman codes. Latest generations of video codecs, including H.264/MPEG-4 AVC and HEVC, employ context-adaptive binary arithmetic coding (CABAC) solely or in combination with other arithmetic coding schemes. This is a lossless compression technique, yet the way it is utilized in video coding standards results a lossy compression.

2.1.2 The Beginning and Early Generations of Video Codecs

In 1984, the International Telegraph and Telephone Consultative Committee (CCITT)[2] published the first generation of video compression standards, H.120. A second revised version was introduced in 1988 that introduced motion compensation and backward prediction. At the time, the codec mainly targeted video conferencing applications but the video quality was not very adequate. Nonetheless, H.120 was a seed that led to much knowledge about video technology. It was not until late 1990 when the ITU-T approved H.261 as a successor to H.120 and first truly practical video coding standard. A second revision was introduced in 1993 adding backward-compatible high-resolution graphics trick mode. H.261 was the first codec to introduce all the major concepts dominating nowadays video codecs such as, 16×16 macroblock motion compensation, 8×8 DCT, scalar quantization, zig-zag scan, run-length, and variable-length coding. These efforts coincided with the ISO/IEC Moving Picture

[2]Now known as the ITU-T Video Coding Experts Group (VCEG), a part of the International Telecommunications Union Telecommunications Standardization Sector (ITU-T), a United Nations organization [2].

Experts Group (MPEG)[3] introducing its first coding standard, MPEG-1 in 1993, which offer a higher bitrate than H.261 and better quality. MPEG-1 inherited the features of H.261 and introduced new tools like bi-directional motion prediction, half-pixel motion, slice-structured coding, DC-only D pictures and quantization weighting matrices. The ITU-T also introduced H.263 coding standard in 1994 as an extension to H.261 and better quality for low bandwidth application on telephony and data networks [27, 28].

In 1994, ITU-T and MPEG jointly developed MPEG-2/H.262 which offered a higher bitrate with support for interlaced-scan pictures and higher DC quantization precision. It also allowed various forms of scalability and concealment of motion vectors for I-frames. MPEG-2/H.262 was widely used for DVD and high-definition video (HDV). It is still used via backward compatibility with following generation of coding standards [23, 27].

The H.263 encoder is based on hybrid DPCM/DCT coding refinements and improvements. H.263 introduced new features and tools including supporting bidirectional MC and sub-QCIF formats. H.263 has several improvements and variations over two versions that were released between 1995 and 2001. These improvements included key enhancements such as error resilience, improved compression efficiency, custom and flexible video formats, scalability for resilience and multipoint, supplemental enhancement information, etc. To facilitate these improvements, techniques like macroblock and block-level reference picture selection, picture header repetition, and spare reference pictures were introduced. Furthermore, the progressive development of H.263 coincided with the introduction of MPEG-4 Part 2, MPEG-4 Visual (formally ISO/IEC 14496-2) which had its first release in early 1999 [2, 27]. MPEG-4 part 2 was based on the H.263 baseline profile adding several new features such as

[3]International Standardization Organization and International Electrotechnical Commission, Joint Technical Committee Number 1, Subcommittee 29, Working Group 11.

increased coding efficiency enhancements, error resilience/packet loss enhancements, segmented coding of shapes, zero-tree wavelet coding of still textures, coding of synthetic and semi-synthetic content, 10 and 12-bit sampling, and others. MPEG-4 part v2 and v3 were later released in early 2000 and 2001, respectively [2, 27].

Following this brief introductory history about the inception and early advances of video coding standards leading to the paradigm-shifting release of H.264/MPEG-4 AVC in 2003, we discuss in the following section the standards that followed since the early 2000s to recent days.

2.1.3 H.264/MPEG-4 Part 10 Advanced Video Coding (AVC)

In 1998, VCEG issued a call for contributions that targeted a new video coding standard that offers double the coding efficiency and capability of MPEG 4 Visual, H.263 or any coding standard in operation at the time. The efforts were mainly led VCEG co-chaired by Gary Sullivan (Microsoft, U.S.) and Thomas Wiegand (Heinrich Hertz Institute, Germany) [11]. By December 2001, MPEG ISO/IEC JTC 1/SC 29/WG 11 joined VCEG to form the Joint Video Coding (JVC) team, which became in charger of finalizing the standard. The final draft and specifications approval eventuated in May 2003, declaring the official introduction of H.264/MPEG-4 Part 10 Advanced Video Coding (AVC). The unprecedented compression efficiency and tools introduced in this standard enabled the revolution of video technology and online streaming applications the world has witnessed over the past decade. In fact, the development of H.264/MPEG-4 AVC continued for over a decade following its initial release which resulted in over 22 amendments, including three major extensions for Fidelity Range Extensions (FRExt), Scalable Video Coding (SVC) and Multiview Video Coding (MVC) [29–32]. To the day of writing this dissertation, H.264/MPEG-4 AVC remains the most dominant video coding standard in active systems with an estimated market share of 74% [33].

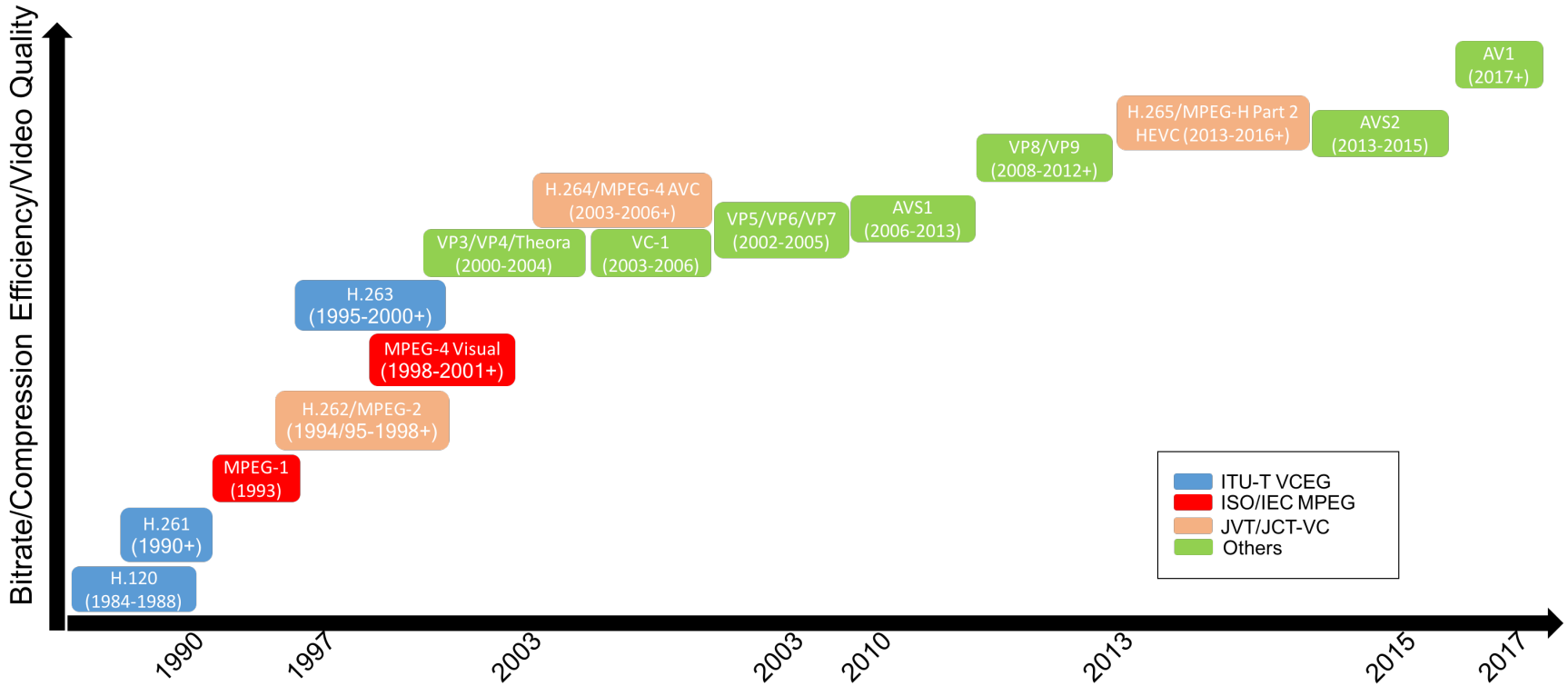


Figure 5: Chronology of video coding standards development and advances [2, 2–10].

The major design enhancements and coding tools introduced in the initial release of H.264/MPEG-4 AVC were mainly in the prediction of picture contents. These new features included [11]:

- Variable block-size motion compensation with small block sizes
- Quarter-sample-accurate motion compensation
- Motion vectors over picture boundaries
- Multiple reference picture motion compensation
- Decoupling of referencing order from display order
- Decoupling of picture representation methods from picture referencing capability
- Weighted prediction
- Improved skipped and direct motion inference
- Directional spatial prediction for intra coding
- In-the-loop deblocking filtering

In addition to prediction, other enhancements and tools in the codec design included:

- Small block-size transform
- Hierarchical block transform
- Short word-length transform
- Exact-match inverse transform
- Arithmetic entropy coding
- Context-adaptive entropy coding

Furthermore, H.264/MPEG-4 AVC included several new features that allowed for more robustness to data errors, networks losses and flexibility for operation over a variety of platforms and networks. These features include:

- Parameter set structure
- NAL unit syntax structure

- Flexible slice size
- Flexible macroblock ordering (FMO)
- Arbitrary slice ordering (ASO)
- Redundant pictures
- Data Partitioning
- SP/SI synchronization/switching pictures

More details and technical aspects of about these features and tools can be found in [11, 27].

2.1.4 H.265/MPEG-H Part 2 High Efficiency Video Coding (HEVC)

The growth and miscellany of video applications combined with the increase in resolution formats and quality beyond high-definition (HD), such as 4k and 8K, were major incentives that motivated various interested parties to investigate coding efficiency superior to H.264/MPEG-4 AVC's capabilities. The ITU-T VCEG and ISO/IEC MPEG began their investigations in mid-2004 and started identifying potential key technology areas (KTAs) to study in early 2005 [34]. A KTA software codebase was developed from H.264/MPEG-4 AVC joint model (JM) to test and verify potential technologies. The investigatory efforts continued in pursuit of technologies that enabled improving the coding efficiency. In January 2010, the two groups formed the joint collaborative team on video coding (JCT-VC) which became the official group overseeing the development and investigation of the project that later was named HEVC. These efforts resulted in the first working draft and first HEVC test model (HM) in October 2010. The development process continued (and still does) until the initial release of the final draft international standard in January 2013. Both ITU-T and MPEG adopted HEVC later that year as approved standards H.265 and MPEG-H Part 2, respectively [11].

HEVC offers double the coding efficiency of AVC maintaining the same picture fidelity. This comes at the expense of high encoding complexity that is estimated

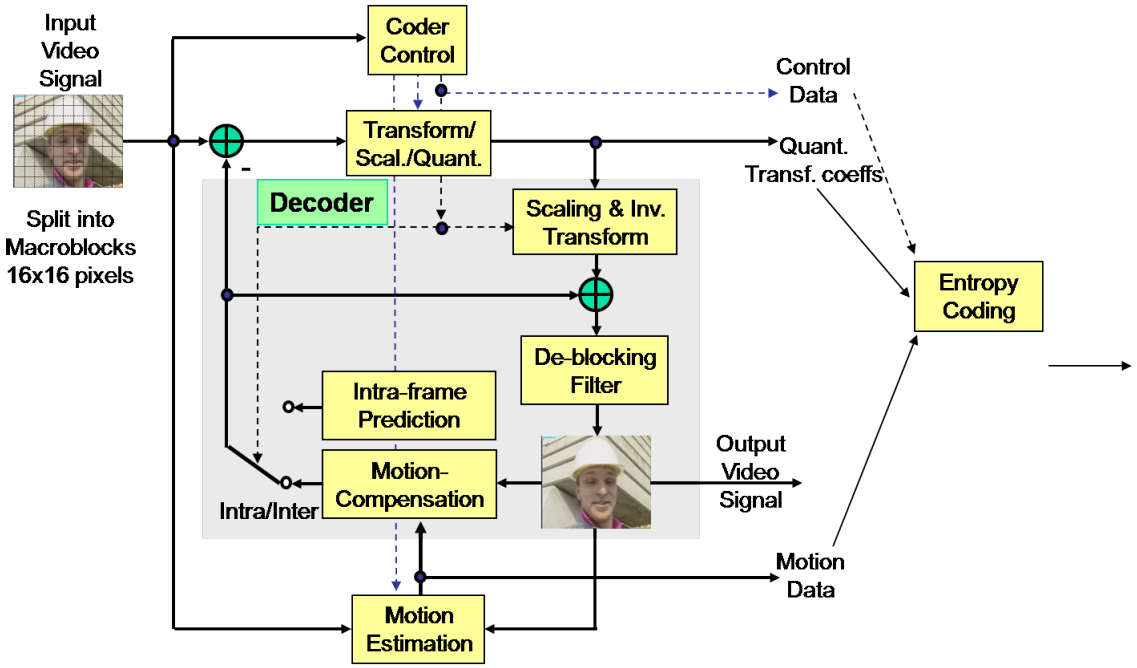


Figure 6: H.264/MPEG-4 AVC video encoder [11].

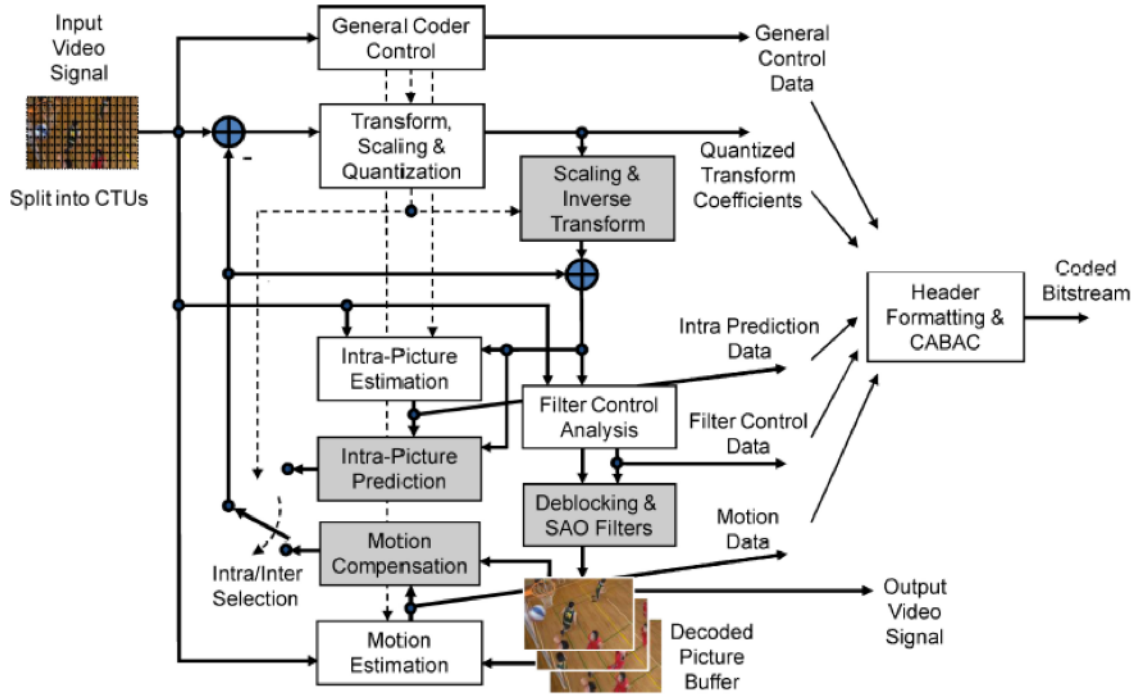


Figure 7: H.265/MPEG-H HEVC video encoder (decoder modeling elements shaded in light gray) [12].

to be ten-fold more than AVC or possibly higher [9, 35]. The coding tools and new features introduced in HEVC exploits tempospatial redundancies more efficiently than its predecessors. This is enabled by higher flexibility in segmentation and block size selection by employing a quadtree structure and signaling. Whereas macroblocks (MBs) are the core of the coding layer in preceding standards, HEVC employs coding tree unit (CTU) and coding tree block (CTB) structure. In this new structure, the encoder has the flexibility to choose CTU size which can be larger than the traditional 16×16 macroblock size supported by AVC. A CTU can be partitioned into additional Coding Units (CU). Furthermore, this flexibility in CTU and CU size is extended to allow segmentation of the luma and chroma components separately using CTBs. Each CTB can be further segmented into coding blocks (CBs) to support more compression efficiency. The same concept is applied to prediction and transform operations by partitioning CBs into Prediction Blocks (PBs) and Transform Blocks (TBs), respectively. Table 1 highlights the major novel technical features introduced in HEVC compared to AVC. Figures 6-7 show the encoder and decoder blocks in both standards [11].

2.1.5 Beyond ITU-T and MPEG: Towards Royalty-Free Video Standards

In the early 2000's, several efforts emerged to develop coding standards and video technology that targeted specific or general applications to compete and possibly replace the standards developed by ITU-T and MPEG. The main motivation behind these efforts was to break a dominant grip MPEG and VCEG had on video technology enforcing substantially high royalty fees charged by companies that hold patents on technologies implemented in these standards [3, 10, 36–38]. As a result, several efforts initiated around the time H.264/MPEG-4 Part 10 AVC was in development and released. We summarize these efforts by categorizing them into the two major directions they reached nowadays without detailing the technical specifications.

Table 1: New features introduced in HEVC compared to AVC.

H.264/MPEG-4 AVC	H.265/MPEG-H HEVC
16×16 Macrobloack	Coding Unit quad-tree structure (64×64 down to 8×8)
Partition down to 4×4	Prediction Unites (64×64 to 4×4) with asymmetric motion prediction
Transforms 8×8 and 4×4 DCT	Transform Units (DCT 32×32 to 4×4 and DST 4×4)
Intra prediction (9 modes)	Intra prediction (35 modes)
Inter prediction luma 6-tap + 2-tap to $\frac{1}{4}$-pel	Inter prediction luma 8-tap to $\frac{1}{2}$-pel and 7-tap to $\frac{1}{4}$-pel
Inter prediction chroma bi-linear interpolation	Inter prediction chroma 4-tap to $\frac{1}{8}$-pel
Motion vector prediction	Advanced motion vector prediction (spatial + temporal)
8b/sample storage & output	8b or 10b/sample storage & output
In-loop deblocking filter	In-loop deblocking filter, Sample Adaptive Offset (SAO)
CABAC or CAVLC	CABAC using parallel operations

Nonetheless, while the basic structure and fundamentals of video coding are common among all these codecs, their performance varies depending on intended application and rigor of development process. Figure 5 shows a chronological progression of the development of these different projects and standards with a comparative performance layout.

Audio Video Coding Standard (AVS): AVS is a working group responsible for digital audiovisual standardization initiated by the government of the People's Republic of China. It was founded by the Ministry of Information Industry in June 2002 and approved by the Standardization Administration of China [5, 36]. This initiative originated from the Chinese government's desire to alleviate the overwhelming royalty and licensing fees China had to pay to foreign companies in order to utilize video technologies developed by MPEG and VCEG. This

working group released the AVS1 standard in 2012. Furthermore, the AVS workgroup released a final committee draft of the second generation, AVS2, in mid 2015. In terms of performance, AVS2 is expected to introduce 50% bitrate saving over previous generations, e.g. H.264 and AVS1, which makes it a competitor to HEVC. Chinese companies own 90% of AVS patents [5]. However, the AVS generation of codecs never gained popularity outside China due to the royalty-free efforts by major industry leaders on the opposite side of the Pacific.

The projects and developments towards open and royalty-free technology took different routes and were driven by several parties in the industry. Most of these efforts have been merging into the recently established Alliance for Open Media (AOMedia) and channeled towards the newly targeted royalty-free video standard, AV1. These efforts can be summarized as follows:

On2 Technologies - TrueMotion VP3-VP8: In May 2000, On2 Technologies[4] introduced its first royalty-free[5] lossy video compression format, On2 TrueMotion VP3. Upon its release, VP3 lacked formal specification for the bitstream format beyond the source code released by On2 Technologies. On2 Technologies released VP4 a year later, which was considered technically complementary to VP3. On2 Technologies continued developing this technology line which resulted in four more releases until the release of TrueMotion VP8 in 2008 [39].

Matroska Multimedia Container: This is a free container and file format based on an open standard developed by the Matroska Development Team and is licensed under GNU L-GPL. It is a file format that can hold an unlimited number of video, audio, picture, or subtitle tracks in one file. The project was

[4]Formerly known as The Duck Corporation.
[5]VP3 was originally a proprietary and patented video codec.

announced in December 2002 and is based on Extensible Binary Meta Language (EBML), a binary derivative of XML. The use of EBML enabled introducing significant advantages in terms of future format extensibility while maintaining file support in old parsers [38].

Xiph.Org Foundation - Theora and Daala: On2 Technologies donated VP3.2 to Xiph.Org Foundation, a non-profit organization that produces free multimedia formats and software tools. The Xiph.Org Foundation started developing Theora in 2001, another free lossy video compression format, which was released in 2004. While the development and releases of Theora continued until its latest release in 2011, another project morphed on its basis in 2004, Daala. Daala's development started in 2004 and continues to this day by Xiph.Org Foundation, Mozilla Corporation, the Internet Engineering Task Force (IETF) and other contributors. Daala had its initial release in May 2013 [40].

Google Inc. - WebM: In February 2010, On2 Technologies was acquired by Google Inc. After the acquisition, VP8 became the core technology to Google's new royalty-free video file format project, WebM, which had its first release in May 2010. After the release of VP9 in December 2012, WebM was updated to support the new standard. A Matroska profile is used to build the WebM container [10].

Cisco Systems, Inc. - Thor: Cisco Systems, Inc. announced in August 2015 the release of their Thor video codec. Thor was denoted to the IETF as well, which has already begun standardization activity towards next generation royalty-free video codec [41, 42]. Nonetheless, Thor has been reported to be very far from being complete and ready for active system deployment [33].

Alliance for Open Media (AOMedia) - AV1: In September 2015, Amazon, Cisco, Google, Intel, Microsoft, Mozilla, and Netflix founded the Alliance for Open Media (AOMedia), a joint development foundation whose soul purpose is the

development of open standards for media codecs and formats amenable to the market and consumers evolving requirements. This move gained popularity on the industry side which prompted more companies to join the alliance including Adobe, AMD, ARM, Broadcom, Nvidia and others. The organization is currently working on the next generation of open, royalty-free video coding format, AOMedia Video 1 (AV1). AV1 is expected to be released in 2017 [3]. All the previously mentioned projects are now channeled towards this technology that promises to fulfill the streaming and market requirement, especially over the web. The Alliance is targeting a 50% improvement over HEVC and VP9 while maintaining reasonable increases in encoding and playback overhead [43]. Given the industrial power and technology behind it, AV1 is expected to be quickly integrated in existing technology and systems, especially the ones provided by contributing companies. To list a few, the Alliance members are primary leaders in the following technologies and services [43]:

- Codec development: Cisco (Thor), Google (VPX), Mozilla (Daala)
- Web browsers: Google (Chrome), Mozilla (Firefox), Microsoft (Edge)
- Content: Amazon (Amazon Video), Google (YouTube), Netflix
- Hardware co-processing: AMD (CPUs, graphics), ARM (SoCs, other chips), Intel (CPUs), NVIDIA (SoC, GPUs)
- Mobile platforms: Google (Android), Microsoft (Windows Phone)
- Over-The-Top (OTT) devices: Amazon (Amazon Fire TV), Google (Chromecast, Android TV)

2.2 Distortions in Videos

In this section, we provide an overview of the different types of perceptual distortions experienced in display and streaming applications. Video distortions can be classified

under four general categories based on the *source or cause of the distortion* [44, 45]:

Acquisition: such as camera noise, motion blur, line/frame jittering, etc.

Compression: these include any artifacts due to lossy compression

Channel-induced: these distortions occur due to video transmission over error-prone networks/channels, such as video freezing, jittering, incorrectly decoded blocks due to erasures, packet loss and delay, etc.

Due to the non-linearity of the quantization process, and the energy distributing effects of the inverse DCT, it is not possible to predict the form or nature of compression distortions or the subsequent channel-induced ones [44, 45].

Post-processing: these occur during post-processing and display, such as post de-blocking and noise

filtering, spatial scaling, retargeting, chromatic aberration, pincushion distortion, etc

In the context of PVQA, video distortions are usually classified into two categories based on the *dimension in pixel domain in which they are perceived*: spatial or temporal distortions. A spatial distortion is perceived or detected in a frame independent of the neighboring frames. In other words, a spatial distortion can be observed or detected in a still image (frame). Temporal distortions represent the temporal inconsistency of spatial features in-between frames. They are identified via the progression, variation, or fluctuation of spatial distortions in pixels, blocks or objects in the temporal domain [45]. Both temporal and spatial distortions negatively impact user's experience. Nonetheless, temporal distortions are more likely to distract human perception and negatively impact quality of experience [45]. The level of distraction also depends on several parameters including distortion magnitude, location of the distortion in the scene, and frame rate.

Yuen and Wu [44] categorized compression and channel-induced distortions as follows:

- Blocking effect
- DCT Basis image effect (quantization)
- Blurring
- Color bleeding
- Staircase effect
- Ringing
- Mosaic Patterns
- False coloring
- False Edge
- MC mismatch
- Mosquito effect
- Temporal fluctuation in stationary area
- Chrominance Mismatch
- Temporal Distortions:
  - Jerkiness
  - Scene Changes
  - Smearing
  - Ghosting
  - up/down-sampling

Furthermore, Zeng et al. [45] classified compression artifacts in recent generation of video codecs as follows:

- spatial:
  - blurring
  - blocking
    - mosaicing effect
    - staircase effect
    - false edge
  - ringing
  - basic pattern effect
  - color bleeding
- temporal:
  - flickering
    - mosquito noise

  - fine-granularity flickering
  - coarse-granularity flickering
- jerkiness
- floating
  - texture floating
  - edge neighborhood floating

2.3 Human Visual System and Visual Perception

In this section, we discuss the nature of the HVS and visual perception features and functions related to video processing and perceptual quality assessment.

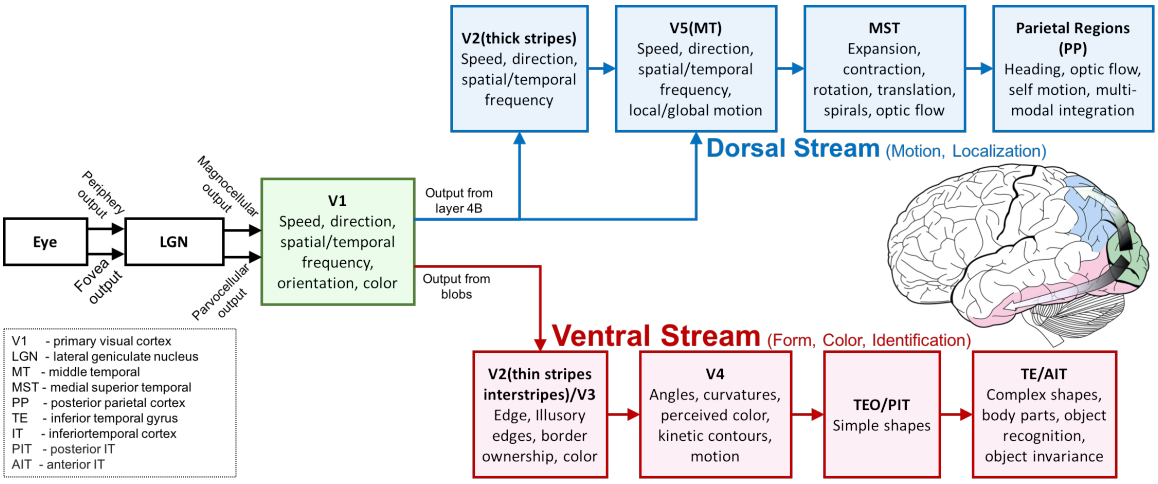


Figure 8: Hierarchical processing of visual streams in HVS (Adopted with changes from [13]).

The process of visual perception begins when light passes through the corneas, pupils, and lenses in our eyes. Light then passes through to the photoreceptors (rods and cones) of the retina where the scene (image) is inversely reflected. The rods (contrast receptors) and cones (color receptors) translates visual data to produce neurological outputs to the ganglion cells in the retina. Two outputs emerge from the ganglion cells to the lateral geniculate nucleus (LGN): (i) M cells outputs and P cells outputs. The M and P cells in the ganglion compliment each other. While M cells are highly sensitive to contrast and lightly sensitive to color, P cells have the opposite sensational properties. Furthermore, M cells receive inputs from both rods

and cones where P cells receive inputs from cones only [13, 46–48].

After passing through the optic nerve and chiasm, this information is passed to the LGN. The LGN performs some spatial correlation and temporal correlations and decorrelations of the visual signals received from both eyes. Visual computations and analysis performed by the LGN serve as a feedback control to the eyes to focus and converge to the principle plane of interest in the object space. The LGN performs stereoscopic and velocity computations determining the relative positions of objects in the visual field associating tags with said objects. This functionality is believed to guide the visual system's attention to important visual information [49].

The majority of signals from LGN are passed to the primary visual cortex, V1, where intricate spatial maps are constructed via rigorous edge-detection and spatial processing focusing on fine spatial and color details. The HVS is believed to operate using two main processing pathways both of which originate from V1, the dorsal stream and ventral stream. Figure 8 shows a hierarchical processing block diagram of the two pathways in the visual cortex. The *ventral stream* is believed to be responsible for object recognition and representation, and color perception [13]. Some studies also associate the ventral stream with long term-memory storage. The *dorsal stream* is believed to be responsible for the motion processing and object locations representation. It is also associated with control of eyes and arms, especially when visual data is required for saccades[6] or reaching control. Figure 8 depicts the various areas in the visual cortex responsible for processing different visual information.

While some details and anatomical aspects of the neurological functions are beyond the scope of this dissertation, we focus here on the characteristics relevant to visual media processing and engineering.

Primary Visual Cortex (V1): This is the earliest and most fundamental area in

[6]Saccade is the term given to rapid, ballistic movements of the eyes that abruptly change the point of fixation [47].

the visual cortex. It excels in spatial processing and pattern recognition. Visual data processing and encoding in V1 is characterized by edge-detection operations. In fact, Gabor filter banks are often used to model the spatial processing performed in V1 [48, 50–52]. The processing of both static scenes and motion begins in V1 and both processing streams (ventral and dorsal) initiate from V1. In fact, studies and recent evidence have shown that this area is inherently multisensory, and not only restricted to visual processing [51, 53].

Vision Field: The HVS perceives the world using two modes of vision: foveal and peripheral vision. Photorecpetors and ganglion cells in the retina are none uniformally distributed. Photorecpetors and ganglion cell are more dense in the the center of the retina (fovea) and gradually become less dense further away from the center (periphery). As a result, foveal vision has the highest sensitivity and visual perception is achieved with the highest resolution and details. Foveal vision primarily targets the point of fixation or center of gaze in the visual field [54]. The resolution of perceived objects decreases gradually as the distance from the fovea (or fixation point) increases. Peripheral vision has a lower perceptual resolution and distinguishable details than foveal vision with a high sensitivity to flicker and motion [55].

Contrast Sensitivity: While the HVS has the ability to perceive a wide range of intensity levels and ambients, the distinction between different brightness values (e.g. shades of gray) is governed by the contrast sensitivity function (CSF). The encoding of brightness (perceived intensity) of an image in the visual cortex relies on local contrast variations. Different intensity levels in an image are distinguished by encoding the variation ratio or difference between intensity levels, instead of the absolute value of intensity. The HVS perceives variations in intensity values within a certain contrast threshold. This means that some

perceived contrast variations in a bright image (scene) may not be perceived by the HVS in a darker one. The minimum perceivable contrast defines the *contrast threshold* and its inverse defines the *contrast sensitivity*. Hence, the CSF defines the perceivable contrast range as a function of spatial frequencies. Utilizing the CSF requires an accurate measurement of local contrasts [54].

Light Adaptation: The HVS has an ability to adapt to different lightness and brightness conditions by controlling the pupils' diameter. This in turn controls the amount of light that passes to the photoreceptors. This nature, in combination with contrast sensitivity, allows the HVS to operate more efficiently over a wide range of intensity values following Weber's law [56].

Spatial and Temporal Masking: The HVS has spatial and temporal masking properties. This refers to the phenomenon of variable ability to detect a visual signal or feature (target) in the presence of another image, scene or context (mask) [57, 58]. This phenomenon is often modeled using just noticeable difference (JND) concept. This is done by estimating the frequency response as a function of the spatial frequency which forms a visual sensitivity model. This model is often used for measuring distortions' visibility and reduce the perceptual effect of compression artifacts. Such models also take into account the CSF of the HVS [58]. The masking effect of the target is stronger when the mask shares the same spatial frequency and orientation. This phenomenon, sometimes characterized by band-pass filtering, explains the well-known property of the HVS that it has a higher sensitivity to distortions in low frequencies than distortions in high frequency bands. Temporal masking can be tougher to characterize and understand. Studies have associated the perception of temporal distortions to visual attention. Furthermore, studies in this domain often assume the HVS to be a limited system [59] which led to more understanding of

the visual contribution to the perception. Nonetheless, several past and recent studies have shown that perception is a multisensory and multimodality process even at the very early stages [46, 48, 51, 53, 60].

Motion Perception: Motion perception start in the V1 area and continues mainly through the dorsal stream (Figure 8). After passing through V2, motion processing takes place in the middle temporal (MT/V5) which also communicates directly with V1. MT cells are believed to be sensitive to properties associated with 2D motion including directionality, speed and spatial frequencies without color selectivity. These characteristics allow the MT area to perform a hierarchal processing of motion fields characterizing both local and global motion patterns. As a result, the HVS can identify local moving points along with the global motion of an object (identified by its edge)[7] [13]. Several other areas in the visual cortex are involved directly or indirectly in motion processing including medial superior temporal (MTS). There has been also evidence that the ventral stream also contributes to motion processing and perception [60]. In general, motion processing is a complicated process in human perception that is relatively less understood than other visual percepts.

2.4 Perceptual Video Quality Assessment

The problem of VQA, in general, has been a popular field of research over the past decade [8-43]. Most of these works target quantifying the video quality in a way that models the HVS's criteria of assessment. There are two ways, in general, to measure the quality of a video: subjective and objective assessment methods.

[7] In Chapter 3, we adopt this approach by applying hierarchical processing of motion fields from gradient-based optical flow maps.

Table 2: A summary of VQA databases.

Database	Year	Resolution	No. of Reference Sequences	Total number of distorted sequences with subjective scores	Frame Rate (fps)	Format	Distortion Types/Hypothetical Reference Circuits (HRC)
VQEG-FR [61]	2000	480i/576i	20	320	25/50	UYVY	H.262/MPEG-2 Part 2, H.263
NYU-PL [62–64]	2007	QVGA	17	12	10-15	YUV	H.264 with packet loss
NYU-Dataset I [65]	2008	CIF/QCIF	6	60	6-30	YUV	Different constant frame rates
IVC-1080i [66]	2008	1080i/VGA/QVGA	24	192		YUV	1080i/VGA/QVGA with ACR and SAMVIQ subjective scores
NYU-Dataset II [67]	2009	CIF	4	68	3.75-30	YUV	Different constant frame rates and QP
IVC-SD RoI [68]	2009	720576	30	84	25	YUV	Normal coding, RoI coding, RoI intra-coding
EPFL-PoliMI [69]	2009	4CIF/CIF	12	156	25/30	YUV, H.264	H.264 with packet loss
LIVE HTTP-based Streaming [70]	2013	720p	3	15	30	YUV	HTTP-based video streaming with varying bitrates
LIVE Mobile [71]	2012	720p	10	200	30	YUV	H.264 compression, wireless packet-loss, frame-freezes, rate-adapted, temporal dynamics
LIVE [72]	2010	768x432	10	150	25/50	YUV	MPEG-2 compression, H.264 compression, H.264 compressed bitstreams with channel-induced distortion
MMSP-SVD [73, 74]	2010	180p, 360p, 720p	3	72	6.5-50	SVC, H.264	SVC (extension of AVC), wavelet-based scalable video coding
MMSP-3D [75]	2010	1080p	6	30	25	AVI	Stereoscopic videos compressed with AVCHD
NYU-Dataset III [76]	2010	CIF	5	180	3.75-30	YUV	Variable and constant frame rates and QP
VQEG-HD [77]	2010	1080p/1080i	9	135	24-30	AVI	15 HRC of H.264 and MPEG-2 cable, satellite, and terrestrial transmission networks and broadband communications services
NYU-Dataset IV [78]	2011	4CIF/CIF/QCIF	7	224	3.75-30	YUV	Variable and constant frame rates and QP
IVP [79]	2011	1920x1088	10	128	25	YUV	DIRAC, H.264 compression, H.264-compressed with error-prone IP networks, and MPEG-2 compression
CSIQ [80]	2014	832x480	12	216	24-60	YUV	MJPEG compression, H.264 compression, HEVC compression, wavelet compression using SNOW codec64, packet-loss wireless networks, and AWGN
IVC-DIBR Videos [81]	2012	XGA	3	102	15-30	AVI	Seven DIBR algorithms used to generate four new viewpoints with different QP's
IVC-Free-Viewpoint synthesized	2014	XGA, 1080p	6	264	15	AVI	2 cameras were choosen to generate 50 intermediate synthesized views
IVC-H264 AVC vs SVC [82]	2010	QVGA, VGA	4	56	30	AVI	Six different H.264 AVC and SVC degredations
IVC-H264 HD vs Upscaling and Interlacing [83]	2010	1080p/1080i	3	87	50	AVI	28 HRC used to compare full-HD format with spatially down-scaled and/or interlaced formats
IVC-Influence Content [84]	2012	VGA	60	300	60	AVI	20 HRC generating different degradations, 4 randomly chosen per sequence
IVC-JEG264HMIX1 [85]	2012	1080p/1080i	10	170	50	AVI	16 HRC of H.264 coding with and without transmission errors, including transcoding scenarios, spatial and temporal subsampling
IVC-NAMA3DS1 COSPAD1 [86]	2012	1080p	10	110	25	AVI	10 HRC for stereoscopic video coding different QPs, bitrates and frame sampling
IVC-SD vs HD H264 [87]	2006	HD, QHD	4	36 (28 HD + 8 SD)	25	AVI	Pair comparison of 2 SD versions vs 7 HD version for every source coded with different H.264 bitrates
IVC-SVC4QoE QP0 QP1 [88]	2011	VGA	11	358	60	AVI	29 HRC with different QPs and configurations of H.264 and H.264/SVC
IVC-SVC4QoE Replace Slice [89]	2010	VGA	9	135	60	AVI	14 different HRCs of h264 and h264/SVC with simulated transmission errors and several error concealments
IVC-SVC4QoE Temporal Switch [90]	2011	VGA	11	424	60	AVI	36 different HRCs of h264 and h264/SVC with Several switching conditions between base and enhancement layer
IVC-VQEG HDTV Pool2 [91]	2010	1080i	9	168	60	AVI, HMIX, PCAP	15 HRCs with different degradations

Subjective quality assessment is considered the best way to establish a ground truth to characterize perceptual video quality. In subjective methods, subjects (viewers) watch video clips and assign a quality score. Average quality for a processed video sequence is known as the Mean Opinion Score (MOS). The latest revision of ITU-T P.910 subjective VQA methods for multimedia applications [14] includes a recommendation of four main categories of methods. First, there is the absolute category rating (ACR) method in which the test sequences are presented one at a time and are rated independently on a category scale. This is a single stimulus method. Secondly, there is the absolute category rating with hidden reference (ACH-HR) method in which the test sequences are presented one at a time and are rated independently on a category scale. The tests must include a reference video of each test sequence, which is displayed as any other test. A differential quality score (DMOS) will be computed between each test sequence and its corresponding (hidden) reference at the analysis stage. This procedure is known as hidden reference. Thirdly, there is the degradation category rating (DCR) method in which the test sequences are presented in pairs. The subject is first presented with the source reference, while the second stimulus is a processed version of the same reference video. This method is a double stimulus impairment scale method. Finally, there is the method of pair comparisons (PC) in which the test sequences are presented in pairs. The same sequence is presented first through one system under test and then through another system. The choice of which method is suitable for a particular application depends on many criteria including context, purpose and the stage of the development process in which the test is performed [14]. Tables 2 summarizes the latest and most common perceptual video quality assessment databases available from different research groups.

The main problem with subjective quality assessment is that it is too demanding in terms of labor and time. Therefore, objective quality metrics are designed to asses

the QoE of the end user. Objective VQA research targets developing quality metrics capable of predicting subjective video quality without subjective tests. A good perceptual objective metric should predict and be statistically well-correlated with subjective data. Since objective video quality evaluation methods are based on mathematical and algorithmic models independent of user feedback, they can be applied in more versatile ways to several applications. Objective quality assessment generally is composed of two steps. First , the image (frame) distortion is estimated producing a local distortion map. The accuracy of the estimation in this step depends on the visual features used and the processing mechanism. Secondly, one pooling function or more are used to statistically merge distortion map values into a single quality score [92]. A detailed discussion of objective quality metrics and their categories is presented in Section 2.5.

Video coding standards generally target improving video streaming efficiency by reconstructing video frames in a compressed manner. However, most codecs, including recent ones, are designed to compress video data by maximizing the peak signal to noise ratio (PSNR) [11, 12]. PSNR is a logarithmic representation of the inverse of mean square error between original and distorted video frames. The higher the PSNR value, the higher fidelity of the distorted frame compared to the original one. Nonetheless, PSNR does not correlate very well with the human perception [93–100]. Hence, there is a need in the both the research and industrial communities to develop accurate ways to estimate video quality as perceived by the HVS. It has been established that the perception of spatial impairments in a video can be reduced in the presence of large motion in the video. Moreover, the perception of temporal distortions is affected by the amount of spatial details [23]. The processing and analysis of video features generally utilizes intensity information because human vision is more sensitive to luminance differences than chromatic ones. Hence, processing and analysis of video features usually takes place on the luminance component of the

video [92, 101]. Furthermore, the HVS is sensitive to contrast variation. Contrast is the deviation between intensity values that makes an object or a region distinguishable. Neurons in the visual cortex of the brain are stimulated by contrast values above a certain threshold [102]. Human contrast sensitivity function determines the visibility of the distortion and level of attention to it [92, 103].

Henceforth, the perceptual quality assessment work in this dissertation takes into account these facts and builds on them. This inspired the design of a new feature space using pixel-level motion fields to address the first step of objective quality metric design. Since the deviation of intensity is important, we do not only quantify these variation in the intensity space. We also take this one step further and track these variations temporally. We utilize optical flow maps at the pixel level to monitor the rate of change in the intensity values in the temporal and spatial domains. The full details of this approach is discussed in Chapter 3.

2.5 Literature Survey

2.5.1 Perceptual VQA Fundamentals

In this section, we review the latest in the field of VQA in general. Objective VQA techniques can be classified into three categories. This classification is based on the magnitude of information required by the VQA technique about the original video sequence. First, there are full-reference (FR) approaches, which require full access to the reference video. These include PSNR, SSIM, NQM, VPM and others. Secondly, there are reduced-reference (RR) approaches, which require a set of coarse features extracted from the reference video sequence to estimate the quality of the reconstructed video. Thirdly, there are no-reference (NR) quality assessment techniques, which do not require any information about the reference video. Furthermore, NR VQA techniques can be divided into three classes. A class of NR approaches, NR-bitstream (NR-B), utilizes only the contents of the received bitstream at the decoder

to estimate the quality of the reconstructed video. Another class of NR approaches, NR-pixel (NR-P), relies only on features and characteristics of decoded frames and pixels to estimate the quality of the reconstructed video. Finally, there are hybrid approaches, NR-bitsream/pixel (NR-BP), which relies on features and information obtained from both the received bitstream and decoded frames. We note here that the framework proposed in this work can best fit under the RR VQA techniques.

The effect of temporal distortion on video quality and the impact of scene motion on perceptual quality is examined in [98]. A subjective quality assessment campaign is conducted using videos with high quality to examine the human response to jerkiness and jitter employing different combinations of strength, duration and distribution of the temporal impairments. The study shows that for low frame-rates, longer impairment duration results in a decreased perceptual quality. Nonetheless, the duration of the impairment is independent of the perceptual quality for high and medium frame rates. Furthermore, reducing the frame rate across the entire video does not cause a significant degradation in perceptual quality. The study also refutes the notion that lower-motion results in better quality compared to higher-motion under the same frame rate degradation. However, head-and-shoulder scenes, a low-motion content, is severely impacted by decimating frame rate.

Seshadrinathan *et al.* [72] conducted a subjective study of video quality on a collection of videos subject to different distortion conditions. The study produced the Mobile LIVE Video Quality Database with 150 distorted videos generated from 10 uncompressed videos. Every video was displayed and evaluated by 38 subjects. The authors also report the performance of several IQA and VQA algorithms on the these videos and correlations with the Difference Mean Opinion Score (DMOS).

In [96], Hemami and Reibman conducted a survey of the relevant work in no-reference quality assessment of images and videos at the time. The study proposes a three-stage paradigm for NR quality assessment allowing for HVS aspects and insights

to be included. The study also benchmarks the different metrics in the study taking into account potential applications and assessment criteria.

The work by Chikkerur *et al.* [97] aims at categorizing the existing metrics based on the features used for quality assessment; natural scenes features or HVS-inspired perceptual features. The authors further subcategorize the metrics in natural scenes class into statistical feauture-based or visual feature-based approaches. Similarly, the metrics in the perceptual class are subcategorized into spatial-domain and frequency-domain approaches. It is reported that the natural visual statistics based MultiScale-Structural SIMilarity index (MS-SSIM), the natural visual feature based Video Quality Metric (VQM), and the perceptual tempospatial frequency-domain based MOtion-based Video Integrity Evaluation (MOVIE) index give the best performance for the LIVE Video Quality Database.

The work of Winkler *et al.* [104] studies, benchmarks and analyzes a large set of image and video quality assessment databases. This work provides a quantitative comparison of these databases in terms of source content, test conditions, and subjective ratings, etc. This kind of works allows researchers to enhance their understanding of these databases and improve on the process of building and designing future databases.

2.5.2 VQA Metrics and Algorithms

Many FR VQA algorithms emerged as extensions to preexisting image quality assessment techniques. Those include PSNR [105], mean square error (MSE), structural similarity index (SSIM) and its variants [93, 106], visual signal to noise ratio (VSNR) [80], and visual information fidelity (VIF) [94]. These FR algorithms operate at the frame level to estimate the video quality using different error and visual criteria. These, however, do not always precisely predict the perceptual quality of videos subject to different types of distortion [95–100].

One of the early standardization attempts for video quality monitoring was the National Telecommunications and Information Administration (NTIA) 2003 standard of the General Model for estimating video quality and its associated calibration techniques. This was the only video quality estimator considered to perform well for both the 525-line and 625-line video tests. In [107], the authors provide a description of the NTIA the General Model and its associated calibration techniques. The study summarizes the test results from the VQEG FRTV Phase II tests and eleven other subjective data sets, all of which were used to develop the method.

LeCallet *et al.* [108] proposed a convolution neural networks (CNN) based objective quality assessment technique to the valuate the perceptual quality of digital videos. This approach uses CNN to continuously evaluate video quality in an attempt to mimic the HVS perception. The paper aims at establishing a foundation for using CNN to combine and pool different objective features extracted from the frames to assess the video quality in a RR configuration. The efficacy of this framework was validated on various MPEG-2 videos with bit rates ranging 2-6 Mb/s. Under these conditions, this approach was reported to correlate well linearly with the recorded subjective scores.

In [109], a *relative* quality metric (rPSNR) is introduced for large-scale, real-time on-line monitoring of streamed video quality. To account for network losses, The authors start by modeling the loss-distortion as a function of application-specific parameters such as codec, error-concealment, bit rate, slicing strategy, scene features, etc. This framework relies on the network to provide a benchmark for quality comparison to facilitate fast real-time quality assessment. This approach was validated by means of simulations and experiments.

In [100], the authors propose a hybrid metrics for video quality estimation in real-time utilizing both bitstream information and pixel features. The study also includes a review of the evolution of video quality metrics and their evolution, and

an overview of the emerging trends in quality measurement at the time. The work in [110] investigates bit rate-distortion and bit rate variability-distortion performance of single-layer video traffic of the H.264/AVC codec and SVC extension. The authors also analyze some frame characteristics from both standards and their impact on traffic and bit rate. The authors report that H.264/AVC codec and SVC extension produce lower bit rate compared to MPEG-4 Part 2. This gain, however, comes at the price of higher traffic variability. The study also examines the effect of this increase in bit rate variability and its impact on frame losses and bufferless statistical multiplexing. The authors show that in some networked applications, the classical way of evaluating the bit rate-distortion improvements is insufficient.

In [111], a hybrid approach utilizing both subjective and objective features to evaluate the QoE of a video over wireless networks is introduced. The proposed Pseudo Subjective Quality Assessment (PSQA) is claimed to minimizes the disadvantages of subjective and objective approaches, which makes it suitable for real-time operations. Naccari *et al.* [21] proposed a no-reference video quality monitoring (NORM) algorithm for assessing the quality degradation with H.264/AVC streamed videos subject to channel induced distortions. The NORM algorithm is a bitstream level quality estimator that operates at the macroblock level at the decoder. It estimates the quality degradation by accounting for effect of error concealment on the spatial and temporal domains, as well as the effect of temporal motion-compensation due to video compression. The authors point out that this approach can be used to predict MOSs in forward prediction systems. The paper includes a RR scenario where this framework is used to predict SSIM values.

In [59], a FR video quality metrics is proposed based on the temporal progression of spatial distortions. This model first evaluates the distortion at eye fixation level leading to a short-term tempospatial pooling of spatial segments. A global score for the whole video is then estimated by performing global long-term pooling of the

Table 3: A summary of the main features used common and recent algorithms and metrics for VQA.

Features \ Algorithm	ST-MAD	ViS$_3$	VRF	3DPoViQ	STAQ	ST-RRED	NORM	PeQASO	Video BLIINDS	VIIDEO	Zhu	Yang	Video CORNIA	Dimitrievski	SACONVIA
	FR				RR				NR						
DCT									■		■			■	
Wavelet					■	■									■
Residuals						■	■		■	■				■	
Motion	■	■		■	■	■	■	■	■	■					
Optical Flow	■	■						■							
Codebook													■		
Deep Network															■
Bitstream							■					■			
Structure				■	■					■			■	■	
3D processing			■	■											■
Distortion specific											■			■	

segmented tempospatial scores. The authors report their approach to have a regular improvement over other video assessment metrics.

The work in [52] introduces a FR VQA motion-based metric, the (MOVIE) index. The MOVIE index takes into account both spatial and temporal (and tempospatial) aspects of distortion assessment. At the time of its development, this metric was reported to compete with algorithms developed and submitted to the VQEG FRTV Phase 1 study in addition to other following metrics. Staelens *et al.* [112] proposed a new subjective assessment methodology for full-length movies. The study highlights the importance of real-life QoE assessment and mainly aims at evaluating the user's audiovisual experience under the same conditions viewers typically watch TV. The authors demonstrate that there exists major differences in terms of impairment visibility and tolerance between the subjective results of their proposed methods, and subjective test conducted using a standardized method.

In [123], the authors investigate the relation between network QoS, application QoS and the user's QoE in HTTP video streaming. The study examines the correlation between network QoS and application QoS both analytically and empirically. The authors then proceed to perform a subjective campaign to determine the relationship between application QoS and QoE. The study reports that variations in the QoE is mainly driven by rebuffering frequency. In [78], the effect of frame rate and quantization on perceptual quality of a video is investigated. The proposed approach evaluates the product of a spatial quality factor and a temporal correlation factor, both of which are modeled analytically. The model's parameters can be estimated from the video features. The authors validate their work on their subjective score and other databases achieving a high Pearson's correlation coefficient with the MOS.

In [124], the technical advancements, standards and proprietary solutions are surveyed in an attempt to define the driving parameters of QoE in HTTP adaptive streaming (HAS) of Internet videos. The study provides a survey of QoE related works from human computer interaction and networking domains. It also revisits the subjective work focused on QoE-driven video adaptation. Consequently, the study identifies the influence factors on QoE and their corresponding models. Furthermore, open issues, conflicting reports and technical factors affecting QoE are discussed. The paper targets researchers and developers concerned with HTTP streaming in general and user-centric QoE. The recent work of Baik *et al.* [125] attempts to quantify the QoE of mobile video streaming. The study considers only three factors as the main of causes of distortion: spatial distortions, types of buffering and resolution changes. Each of these are modeled using machine learning to estimate their contribution to quality degradation. This is network-level approach without including any visual features, image/video processing or HVS insights.

In this dissertation, we address the problem of PVQA by focusing on pixel-based techniques. Pixel-based approaches are more accurate since they examine the visual

stimuli that viewers observe. They also do not rely on the bitstream metadata which is insufficient and may not be accessible. We design algorithms and metrics utilizing novel features that have not been fully investigated in previous work. Namely, we design a metric utilizing pixel-level motion fields. Furthermore, we also propose two metrics utilizing the tempospatial characteristics of the power spectral density of video contents. The proposed metrics incorporate both spatial and temporal features of the video. In addition, we also analyze different visual feature maps and their corresponding temporal and spatial statistical moments to determine their correlation with perception. This analysis improves our understand of human perception and its sensitivity to distortion. This knowledge is essential in the design and verification of accurate perceptual quality metrics.

Table 4: A summary of common and recent algorithms and metrics for VQA.

Metric/Model	Year	Key Features/Approach	Reported Testing/Validation/Databases
PSNR [105]	-	log of the inverse of MSE	
NQM [126]	2000	variation in contrast sensitivity/masking with distance image dimensions, and spatial frequency	static images
VSNR [127]	2002	pixel-based visual masking and aggregation	LIVE image DB
SSIM, MS-SSIM [93]	2003	natural scene stats, structural similarity	VQEG Phase I, LIVE image DB
VQM [107]	2004	natural scene feature, edge impairment filter	VQEG FRTV Phase II
VIF [94]	2006	natural scene stats, wavelet domain visual model of HVS	static images with different distortions
Le Callet et al. [108]	2006	HVS modeling using CNN with variant features	MPEG-2 videos, with bit rates ranging 2-6 Mb/s
rPSNRn [109]	2008	uses network feedback to estimate channel-induced distortion independent of content	simulations using CIF/QCIF videos
Winkler and Mohandas [100]	2008	pixel/bitstream based fidelity assessment	simulation based validation with network information
NORM [21]	2009	analytical estimation of distortion using packet loss information and error propagation estimation	Comparison with subjective scores of SD, CIF and QCIF videos
Piamrat et al. [111]	2009	objective metric using RNN to model quality behavior and predict MOS	Subjective tests of simulated scenarios and compare with PSNR
Ninassi et al. [59]	2009	uses eye fixation level to decide tempospatial pooling of spatial	simulated scenarios and comparison with PSQA and PSNR
MOVIE [72]	2010	Frequency doamin, banks of Gabor filters	VQEG FRTV Phase I
Brandao and Queluz [128]	2010	transform domain, bitstream-based, DCT coefficients to estimate PSNR	11 CIF H2.64 coded videos with MOS correlations
Ou et al. [78]	2011	Parametric modeling of spatial and temporal quality using video features	CIF/QCIF videos with different coding parameters and MOS
TQV [129]	2012	uses ML as to determine weight of temporal/spatial parameters in overall VQA	EPFL-PoliMI DB, LIVE video DB, TUL DB
ViS3 [15]	2014	adaptively estiamte spatial distortion across frame followed by joint tempospatial calculation	Mobile LIVE DB, IVP DB, CSIQ DB
Baik et al. [125]	2015	bitstream-based, ML modeling of spatial, buffering and resolution change distortions	large pool of online videos subjectively evaluated
PeQASO	2015	Pixel-based spatial, temporal tracking of intensity inconsistency due to distortion using optical flow	Mobile LIVE DB, IVP DB, CSIQ DB

CHAPTER III

PERCEPTUAL QUALITY ASSESSMENT USING OPTICAL FLOW FEATURES (PEQASO)

The research community has been investigating different criteria and features in the digital video processing to model the assessment methodology of the HVS [21, 58, 78, 108, 111, 124, 129–134]. As it was explained earlier, perceptual quality assessment highly depends on the human sensitivity to contrasts and brightness variations in both the spatial and temporal domains. Furthermore, video codecs use block-based operations to temporally compress the video signal. As the survey of prior art showed, VQA algorithms usually examine block-based motion vectors (and sometimes the residuals) to analyze the temporal features of the video. We argue here that these compressed descriptors do not provide sufficient information about the distortion or video dynamics. Furthermore, these elements ignore the human visual criteria of perceptual assessment, which is the sensitivity to intensity and contrast. Since the deviation of intensity is important from a human perception standpoint, we quantify these variations in intensity and track these variations temporally. Thus, we propose utilizing gradient-based optical flow at the pixel level as an intensity variability map. By examining and processing the optical flow maps at various scales, we can estimate the distortion within a frame at the pixel level by capturing the inconsistencies in these optical flow maps. This approach incorporates the hierarchal local and global processing mechanism of motion in the visual system alluded in Section 2.3 [131–133].

3.1 Optical Flow Preliminaries

There are several approaches to computing motion fields in videos. Block-matching is the most commonly used approach, especially in video coding [11, 12, 23]. Motion fields generated using block-based approaches do not capture all the details and variations at the pixel level. Thus, we focus here on gradient-based estimation, which produces a pixel-level motion map [135–137]. Optical flow methods estimate the motion between two frames taken at times t and $t + \Delta t$ at every pixel. Assuming that pixel intensities are translated from one frame to the next, the intensity value at location (x, y, t), $I(x, y, t)$, will change by Δx, Δy and Δt between two frames. Assuming the intensity to be constant along a motion trajectory, this implies the following constraint [27, 138]:

$$\frac{dI(x, y, t)}{dt} = 0 \tag{1}$$

This expression denotes the rate of change in intensity along the motion trajectory. Using the chain rule, the constraint in (1) can be approximated as follows:

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0 \tag{2}$$

$$\frac{\partial I}{\partial x}U + \frac{\partial I}{\partial y}V + \frac{\partial I}{\partial t} = 0 \tag{3}$$

where $U = \frac{\Delta x}{\Delta t}$, $V = \frac{\Delta y}{\Delta t}$ are the x and y components of the velocity or optical flow of $I(x, y, t)$ and $\frac{\partial I}{\partial x}$, $\frac{\partial I}{\partial y}$ and $\frac{\partial I}{\partial t}$ are the derivatives of the frame at (x, y, t) in the corresponding directions. This expression is known as the optical flow equation (OFE). Since the equation contains two variables, it cannot be solved directly. This is the aperture problem of optical flow estimation [27, 138]. Optical flow methods introduce additional constraints for estimating the actual flow. The gradient-based approach in [135] for instance makes use of finite differences to overcome the aperture problem. However, the proposed VQA framework does not rely on the calculation method and can be deployed using any optical flow algorithm.

Let the distortion in the frame at location (x, y, t) be denoted by $\epsilon\,(x, y, t)$. The intensity value in the distorted frame $I\,(x, y, t) + \epsilon\,(x, y, t)$, where $0 \leq I\,(x, y, t) + \epsilon\,(x, y, t) \leq 1$. Hence, the expression in (3) can be written as follows:

$$\frac{\partial I}{\partial x}U + \frac{\partial I}{\partial y}V + \frac{\partial I}{\partial t} + \frac{\partial \epsilon}{\partial x}\mu + \frac{\partial \epsilon}{\partial y}\nu + \frac{\partial \epsilon}{\partial t} = 0 \tag{4}$$

where $\mu = \frac{\Delta x}{\Delta t}$, $\nu = \frac{\Delta y}{\Delta t}$ are the x and y components of the velocity or optical flow of $\epsilon(x, y, t)$ and $\frac{\partial \epsilon}{\partial x}$, $\frac{\partial \epsilon}{\partial y}$ and $\frac{\partial \epsilon}{\partial t}$ are the derivatives of the distortion signal in the frame at (x, y, t) in the corresponding directions. Furthermore, let $\rho(x, y, t)$ denote the magnitude of the distortion signal flow velocity,

$$\rho(x, y, t) = \sqrt{\mu(x, y, t)^2 + \nu(x, y, t)^2}. \tag{5}$$

The magnitude $\rho(x, y, t)$ represents the rate of change in the intensity signal causing a distortion between frames. Hence, the aggregate of this signal in the spatial domain, $\sum_{\forall x} \sum_{\forall y} \rho(x, y, t)$, yields a quantification of motion field causing the distortion at any instant of time, t. In the following section of this chapter, we propose an algorithm to process the optical flow maps to estimate the motion field causing a perceptual distortion in the frame. We process the optical flow map at multiple spatial scales and estimate a multi-scale perceptual distortion moving from local to more global descriptors in the frame.

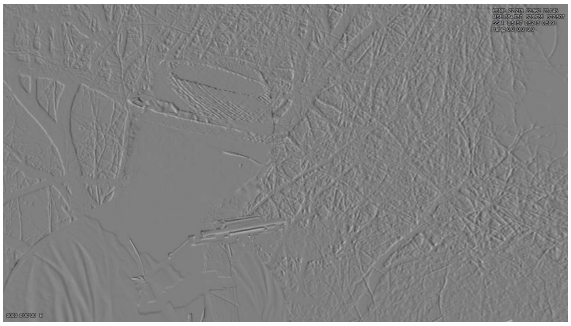
3.1.1 Video Quality Monitoring Using Optical Flow

Figure 9 shows the visualizations of the optical flow maps of original frames, decoded frames with compression artifacts only, and frames with channel-induced distortion, respectively[1]. The optical flow tends to be less homogeneous with the increase of distortion in the frames. These optical flow images clearly show that any distortion or artifact in the frame will cause a disturbance to the original natural motion field. Furthermore, Figure 10 shows empirically estimated probability density functions

[1]The optical flow images in Fig 9 were generated using the software implementation in [139].

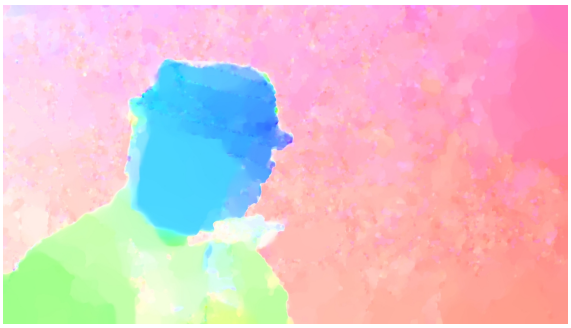
(a) Anchor frame 240

(b) Anchor frame 241

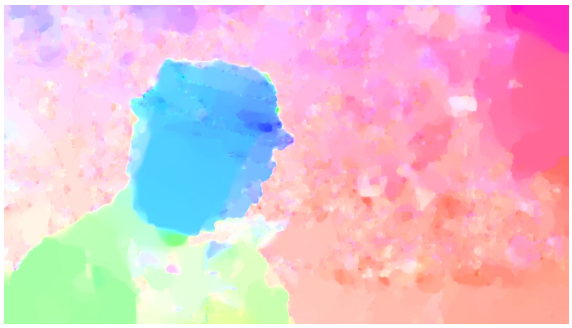


(c) Residual frame

(d) Anchor optical flow



(e) Optical flow of frame with compression artifacts only



(f) Optical flow of frame with channel-induced distortion

Figure 9: Example from video `HC` in the Mobile LIVE database [28] that shows the differences between the optical flow maps of the anchor, error-free decoded and distorted frames.

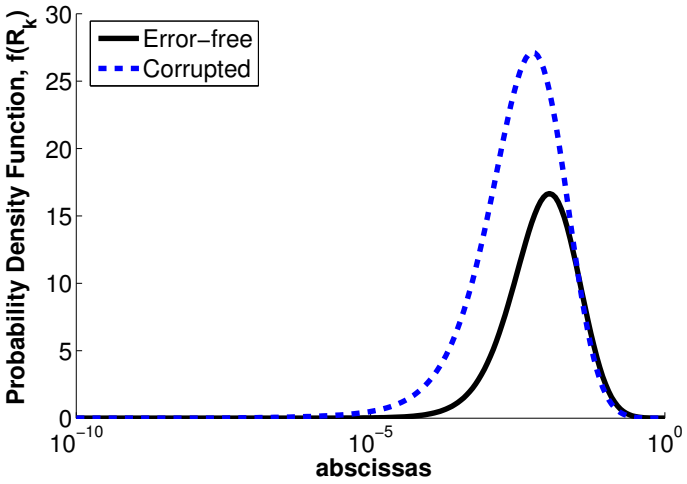


Figure 10: Probability density function of the optical flow maps in Figure 9 (e) and (f).

(PDFs) of these optical flows, $\mathbf{R}_k$, for the error-free and corrupted frames. The figures show that there is a discrepancy in the optical flow caused by distortion in the frame. The change in the PDF of the optical flow map shows that the statistical features of the error-free map are different from the statistics of the distorted map. It should be noted that the SSIM values of the corrupted frame in this case is 0.998.

In this work, our goal is to capture these inconsistencies in the optical flows throughout the video due to the channel-induced distortions. Let f_k be the frame of interest. Furthermore, let $\mathbf{U}_k$ and $\mathbf{V}_k$ denote the matrices of its horizontal and vertical optical flow velocities, respectively. Furthermore, let $\mathbf{R}_k$ denote the matrix of magnitudes of the flow velocities [135]:

$$\mathbf{R}_k = \sqrt{\mathbf{U}_k^2 + \mathbf{V}_k^2}, \tag{6}$$

where k is the temporal index of the frame in the received video. If the spatial dimensions of the frame are $M \times N$, then the dimensions of $\mathbf{U}_k$, $\mathbf{V}_k$, and $\mathbf{R}_k$ are also $M \times N$. All the results and experiments in this proposal were obtained using the Horn-Schunck optical flow estimation method [135]. Nevertheless, the processing framework introduced herein is valid for any optical flow estimation algorithm.

Figure 11 shows a flow chart of the process performed to each optical flow map

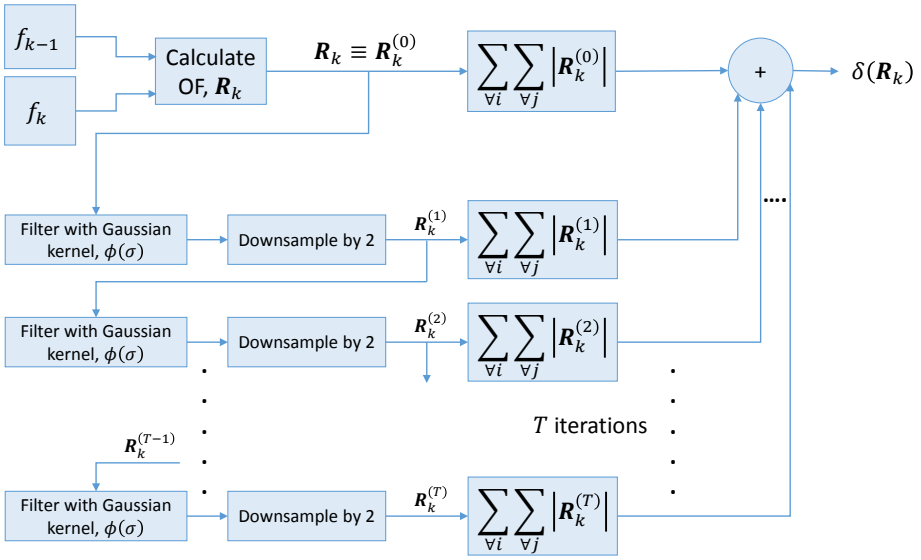


Figure 11: Iterative optical flow frame-level processing flowchart.

of every frame. In an iterative manner, we examine the aggregate of magnitudes of the optical flow maps at different scales. This approach was inspired by the diffusion distance dissimilarity metric [140]. The following expressions are the core of this iterative process:

$$\mathbf{R}_k^{(l)} = \text{Downsample}_2\left[\mathbf{R}_k^{(l-1)} * \phi(\sigma)\right] \text{ and} \tag{7a}$$

$$\delta(\mathbf{R}_k) = \sum_{l=0}^{T}\sum_{\forall i}\sum_{\forall j}\left|R_k^{(l)}(i,j)\right|, \tag{7b}$$

where $\mathbf{R}_k^{(l)}$ is the downsampled version of $\mathbf{R}_k^{(l-1)}$ in iteration l, T is the total number of iterations, $\left|R_k^{(l)}(i,j)\right|$ is the magnitude of the optical flow at pixel (i,j), and $\phi(\sigma)$ is a Gaussian kernel with standard deviation σ. In every iteration, we downsample after smoothing the map with a Gaussian kernel. Then, we take the aggregate of the magnitudes of this filtered version of optical flow map as an output from each iteration. The number or iterations in this study was fixed to three iterations. The aggregates from all the iterations are finally accumulated into one descriptor. This process yields a descriptor of an optical flow map, $\delta(\mathbf{R}_k)$.

Assuming a RR quality estimation configuration, the reference descriptor $\delta(\mathbf{R}_{k,\text{ref}})$,

is estimated at the encoder and available at the decoder or quality estimator. The decoder performs an identical operation and calculates the descriptor for the received frame, $\delta\left(\mathbf{R}_{k,\mathrm{rx}}\right)$. The difference between the descriptors from the reference and received videos are then used as a perceptual quality estimator as follows:

$$D_k = \left| \log \left[\frac{\delta\left(\mathbf{R}_{k,\mathrm{rx}}\right)}{\delta\left(\mathbf{R}_{k,\mathrm{ref}}\right)} \right] \right|. \tag{8}$$

We denote D_k as the estimated distortion in frame k. This captures the inconsistencies in the optical flow map at the frame level. Following the notation and optical flow formulation alluded earlier, our RR algorithm quantifies the aggregates of the matrices:

$$\frac{\sum\sum\left(\mathbf{R}_k + \rho_k\right) + \sum\sum\left(\mathbf{R}_k^{(1)} + \rho_k^{(1)}\right) + \sum\sum\left(\mathbf{R}_k^{(2)} + \rho_k^{(2)}\right) + \cdots}{\sum\sum\mathbf{R}_k + \sum\sum\mathbf{R}_k^{(1)} + \sum\sum\mathbf{R}_k^{(2)} + \cdots} \tag{9}$$

at multiple spatial scales to measure a frame's distortion moving from local to global descriptors. Hence, this metric captures the inconsistencies in the optical flow map at the frame level. To estimate the perceptual quality at the sequence or GOP level, $\mathcal{P}$, we simply calculate the arithmetic mean for the picture set of interest:

$$\mathcal{P} = \left[\underset{\forall k}{\mathrm{E}} \left[D_k \right] \right]^{\alpha} \tag{10}$$

where α is an empirically determined sequence-dependent parameter.

3.2 Experiments and Results

In this section, we discuss the experimental results and validation for the proposed perceptual quality estimation framework. We tested the proposed framework on a variety of test sequences subject to channel-induced distortion. The test sequences were selected from three independent video quality assessment databases [15, 71, 79]. We show the results of each of these databases independently and the overall accuracy of the perceptual quality estimation across databases.

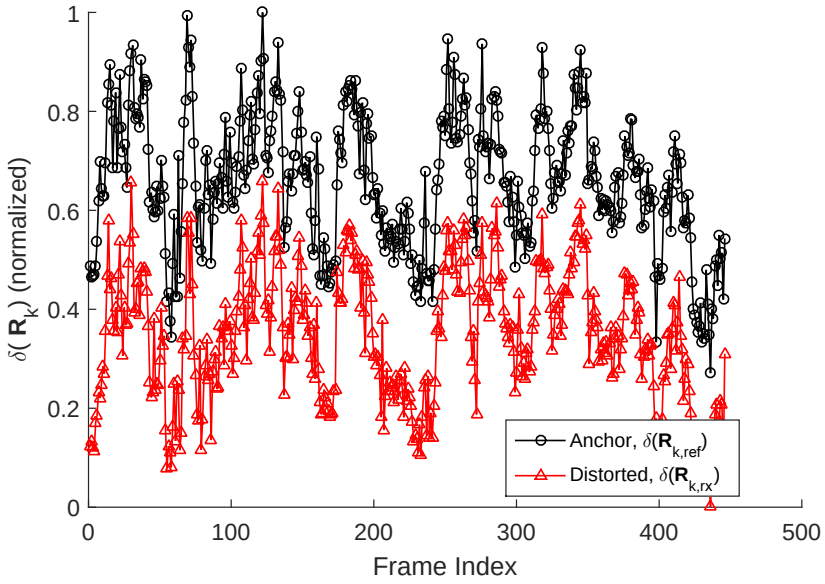


Figure 12: A sample of the difference between $\delta(\mathbf{R}_{k,\mathrm{rx}})$ and $\delta(\mathbf{R}_{k,\mathrm{ref}})$ for `HC` sequence with H.264 compressed video with channel-induced distortion.

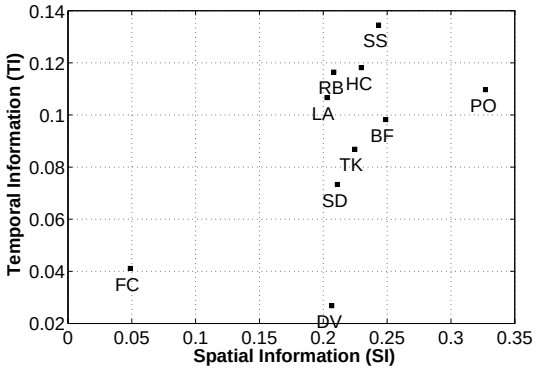
3.2.1 Video Quality Assessment Databases

3.2.1.1 Mobile LIVE Video Quality Assessment Database [71]

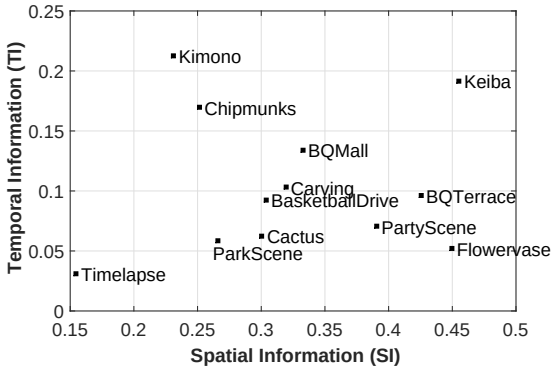
The Mobile LIVE video database was developed at the University of Texas at Austin. It contains 10 reference videos. There is a total of 40 distorted videos with H.264 compression artifacts and 40 with channel-induced distortion. The videos were compressed using H.264 SVC and transmitted over a simulated wireless IEEE 802.11 channel. To simulate channel distortion, these transmissions were subject to packet loss to degrade the perceptual quality. Both reference and distorted videos were provided in raw `YUV420` format with a resolution of 1280×720 pixels. The duration of these videos is 10 s at frame rates of 25 fps.

3.2.1.2 CSIQ Video-Quality Database [15]

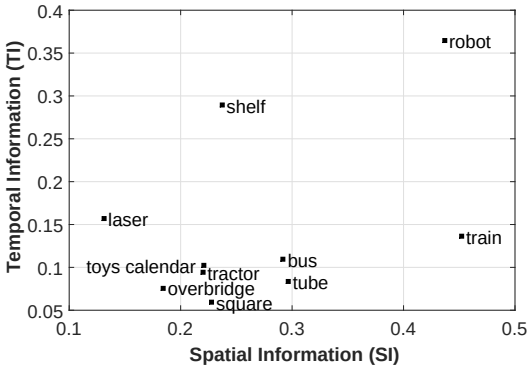
The CSIQ video-quality database was developed at Oklahoma State University Vision Lab. It contains 12 reference videos, 34 with H.264 compression distortion and 34 distorted videos with channel-induced distortion. Videos were compressed using



(a) Mobile LIVE Video Quality Assessment Database [71]



(b) CSIQ Video Da5tabase [15]



(c) IVP Subjective Quality Video Database [79]

Figure 13: Temporal Information (TI) and Spatial Information (SI) indices for the three evaluated databases.

H.264 SVC and transmitted over a simulated wireless channel. These transmissions were subject to packet loss to introduce distortions degrading the perceptual quality. Both reference and distorted videos are provided in raw `YUV420` format with a resolution of 832×480 pixels. The duration of these videos is 10 s at frame rates ranging from 24 to 60 fps.

3.2.1.3 IVP Subjective Quality Video Database [79]

The IVP subjective quality video database was developed at the Chinese University of Hong Kong. It contains 10 reference videos, 40 with H.264 compression, and 28 distorted videos with channel-induced distortion. The distorted videos were obtained

via simulated transmission of H.264-compressed bitstreams through error-prone IP networks. These transmissions were subject to packet-loss to introduce distortions degrading the perceptual quality. Both reference and distorted videos are provided in raw `YUV420` format with a resolution of 1920×1088 pixels. The duration of these videos is 10 s at a frame rate of 25 fps.

Figure 13 temporal information (TI) and spatial information (SI) indices for the three evaluated databases. These plots where obtained using the P.910 subjective video quality assessment recommendation of the TU Telecommunication Standardization Sector [14]. These plots show that the Mobile LIVE database mostly contains sequences of low temporal and spatial complexity. The CSIQ video quality database is mostly clustered in a medium to high spatial complexity and medium to low temporal complexity. Finally, the IVP subjective quality database mostly contains videos of medium spatial complexity and a widely varying range of temporal complexity.

3.2.2 Performance Metrics and Auxiliary Formulation

In this section, we briefly describe the performance metrics used in evaluating perceived video quality estimators and metrics performance.

3.2.2.1 Linearity

Pearson correlation coefficient is used to measure the linearity of the predictions which is formulated as

$$PLOCC = \frac{\sum_{s=1}^{T}(x_s - \mu_x)(y_s - \mu_y)}{\sqrt{\sum_{s=1}^{T}(x_s - \mu_x)^2} \cdot \sqrt{\sum_{s=1}^{T}(y_s - \mu_y)^2}}, \tag{11}$$

where x_s is the estimated score and y_s is the mean opinion score corresponding to a video indexed with s, μ is the average operator, and T is the total number of videos.

3.2.2.2 Ranking

Spearman correlation coefficient is used to measure the monotonic relationship between quality estimates and subjective scores. Instead of using exact values, ranks

of the values are used. For example, let the total number of videos be T with corresponding mean opinion scores (y_s). Based on the rankings, the minimum score should be ranked as 1, the maximum as T, and the others should be in between 1 and T based on their rankings. This process is applied to both subjective scores and estimates. If the relative order of the subjective scores and the objective estimates are same, correlation should be 1.0 otherwise it should be lower. The formulation of Spearman correlation coefficient is given as

$$SROCC = 1 - \frac{6 \sum_{s=1}^{T} (X_s - Y_s)^2}{T \cdot (T^2 - 1)}, \tag{12}$$

where X_s is the rank assigned to the score x_s and Y_s is the rank assigned to the subjective score y_s, which corresponds to video indexed with s, and T is the total number of videos.

Kendall rank correlation coefficient is also based on ranking but we do not directly assign rankings to all estimates and scores. Instead, estimates and scores are compared one by one. For example, x_s is the estimate and y_s is the mean subjective score corresponding to a video indexed with s and we have x_l and y_l corresponding to an video indexed with l. If $x_s > x_l$ and $y_s > y_l$ or $x_s < x_l$ and $y_s < y_l$, these pairs are denoted as `concordant`. If $x_s > x_l$ and $y_s < y_l$ or $x_s < x_l$ and $y_s > y_l$, these pairs are denoted as `discordant`. Finally, if $x_s = x_l$ and $y_s = y_l$, this pair is neither concordant nor disconcordant. Once all of the pair combinations are considered, Kendall correlation coefficient is calculated as

$$KROCC = \frac{(T_{cor}) - (T_{dis})}{0.5 \cdot T \cdot (T - 1))}, \tag{13}$$

where T_{cor} is the number of concordant pairs, T_{dis} is the number of discordant pairs, and T is the number of videos in a set.

3.2.3 Results and Validation

Figure 14 show scatter plots of the perceptual quality predictor, $\mathcal{P}$, versus the reported DMOS scores in the databases. Fig 14a shows a scatter for sequences with

H.264 compression artifacts, while Figure 14b is a scatter plot for sequences with channel-induced distortion. The plots show a good linear correlation of the tested decoded sequences from the three databases. While this approach is pixel-based and not limited by the codec, parameters or configuration, the results shows herein are based on H.264 compressed videos sequences only due to their availability in all the databases. Furthermore, Tables 5-6 report Spearman's rank-order correlation coefficient (SROCC) and Pearson-linear order correlation coefficient (PLOCC) for the three databases separately. These statistics are provided for PeQASO and nine other FR quality popular and well-accredited video and image quality assessment metrics. The proposed work in [21] is the closest RR approach to our work for comparison purposes. However, this algorithm operates on the bitstream metadata and the implementation was done for older versions of H.264 bitstreams format. Furthermore, recent databases provide only YUV video files without bitstreams. More importantly, the algorithm is designed to estimate channel-induced distortion by examining metadata of the bitstream without taking into account compression artifacts. Hence, the comparison with the NORM algorithm would not provide an objective assessment of PeQASO's performance. Therefore, it was not possible to compare with similar RR quality assessment metrics. The most feasible solution was to compare the performance with FR metrics where the reference for all the videos is the anchor video without any distortion. We note here that the comparison with FR reference metrics represents a more challenging competition for our RR metric given that the amount of data required and computational complexity are much higher in FR metrics.

3.2.3.1 H.264 Compression Artifacts

We can observe that PeQASO is competitive with all the reported FR metrics. In terms of SROCC, our metric ranked first on the mobile LIVE and IVP databases. It also ranked second on the CSIQ with a marginal depreciation of 0.005 from the

Table 5: SROCC and PLOCC for the proposed metric and a set of the most common video and images quality assessment metrics tested on sequences with H.264 compression artifacts only.

VQA Metric	PSNR [105]	VQM [107]	MOVIE [72]	TQV [129]	ViS3 [15]	MS-SSIM [93]	VIF [94]	VSNR [127]	NQM [126]	PeQASO (Proposed)	ST-RRED [113]	Video BLIINDS [114]	VIIDEO [115]
Reference	Full									Reduced		No-Reference	
SROCC													
LIVE	0.819	0.772	0.7738	0.764	0.757	0.804	0.861	0.874	0.850	**0.975**	0.956	0.692	0.242
CSIQ	0.802	0.919	0.897	0.955	0.920	0.679	**0.984**	0.637	0.773	0.979	0.977	0.341	0.625
IVP	0.866	0.862	0.823	0.672	0.876	0.687	0.903	0.771	0.733	**0.908**	0.861	0.058	0.11
PLOCC													
LIVE	0.784	0.782	0.8103	0.788	0.773	0.766	0.883	0.849	0.832	**0.922**	0.797	0.676	0.275
CSIQ	0.835	0.916	0.904	0.965	0.918	0.658	**0.989**	0.607	0.779	0.981	0.874	0.301	0.624
IVP	0.855	0.869	0.744	0.898	0.898	0.662	0.909	0.735	0.746	**0.924**	0.589	0.035	0.154

Table 6: SROCC and PLOCC for the proposed metric and a set of the most common video and images quality assessment metrics tested on sequences with channel-induced distortion.

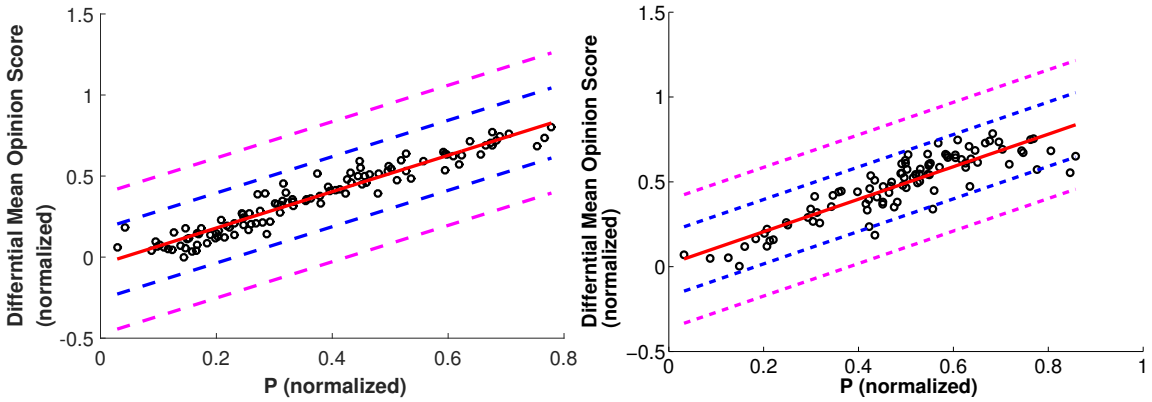
VQA Metric	PSNR [105]	VQM [107]	MOVIE [72]	TQV [129]	ViS3 [15]	MS-SSIM [93]	VIF [94]	VSNR [127]	NQM [126]	PeQASO (Proposed)	ST-RRED [113]	Video BLIINDS [114]	VIIDEO [115]
Reference	Full									Reduced		No-Reference	
SROCC													
LIVE	0.793	0.776	0.6508	0.754	0.845	0.813	0.874	0.856	0.899	**0.972**	0.969	0.654	0.28
CSIQ	0.851	0.801	0.886	0.842	0.856	0.757	**0.954**	0.889	0.866	0.803	0.848	0.089	0.007
IVP	0.711	0.650	**0.858**	0.629	0.807	0.620	0.690	0.766	0.795	0.727	0.665	0.367	0.129
PLOCC													
LIVE	0.762	0.791	0.7266	0.777	0.846	0.771	0.898	0.849	0.874	**0.948**	0.831	0.636	0.312
CSIQ	0.802	0.806	0.882	0.784	0.850	0.761	**0.963**	0.899	0.889	0.841	0.513	0.115	0.015
IVP	0.673	0.642	**0.842**	0.735	0.802	0.638	0.633	0.721	0.736	0.824	0.292	0.172	0.227

Table 7: Statistical correlation of the proposed metric for different temporal and spatial features for sequences with H.264 compression artifacts.

Region	temporal complexity	spatial complexity	No. of reference videos	Total no. of distorted videos	SROCC	PLOCC	KROCC
R_1 (top)	high	variant	10	36	0.882	0.929	0.787
R_2 (bottom left)	low	low	11	43	0.869	0.863	0.665
R_3 (bottom right)	low	high	10	33	0.804	0.819	0.625
All	variant	variant	31	112	0.968	0.965	0.841

Table 8: Statistical correlation of the proposed metric for different temporal and spatial features for sequences with channel-induced distortion.

Region	temporal complexity	spatial complexity	No. of reference videos	Total no. of distorted videos	SROCC	PLOCC	KROCC
R_1 (top)	high	variant	9	31	0.925	0.917	0.802
R_2 (bottom left)	low	low	10	39	0.817	0.750	0.622
R_3 (bottom right)	low	high	9	26	0.811	0.822	0.607
All	variant	variant	28	96	0.898	0.890	0.719



(a) Sequences with H.264 compression atricafts. (b) Sequences with channel-induced distortion

Figure 14: Scatter plot of the perceptual quality predictor, P, versus the reported DMOS in the three subjective quality video databases. The blue and pink lines are $P \pm \sigma$ and $P \pm 2\sigma$, respectively, where σ is the data standard deviation. The red line is the mean.

best performing metrics. Similarly, in terms of PLOCC, our metric ranked first on both Mobile Live and IVP. In the CSIQ database, it again ranked second with a depreciation of 0.008 from the best performing metric. This shows that PeQASO is very suitable for online perceptual evaluation of a wide variety of sequences spanning different spatial and temporal features.

We also tested the proposed metric across the databases while fixing the range of temporal and spatial complexity range. The goal of this test is to benchmark the metric in light of the video features eliminating any database-based bias. Figure 15 shows three different regions where we pooled the sequences based on their temporal and spatial features regardless of the database. We note here that the videos across the databases vary in coding configurations and parameters, nature of channel errors, resolution, etc. The division of the three regions in Figure 15 was based on pooling sequences with similar features together. We also tried to provide sufficient numerical data for statistical analysis in every region. The top devision, R_1, includes highly temporally complex videos with varying spatial information complexity. The bottom left devision, R_2, contains low temporally complex videos with low spatial information

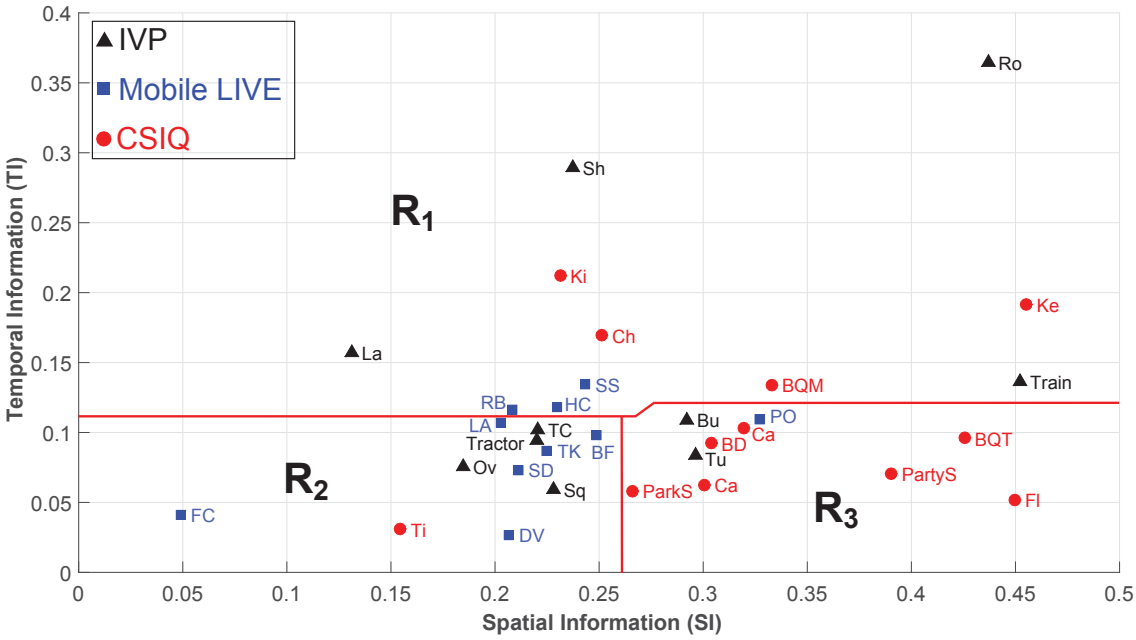


Figure 15: Temporal Information (TI) and Spatial Information (SI) indices for the all the tested sequences across the three databases.

complexity. The bottom right devision, R_3, contains low temporally complex videos with high spatial information complexity.

The correlation coefficients of this test are shown in Table 8. In addition to SROCC and PLOCC, we also report Kindell's rank-order correlation coefficient (KROCC). The results in Table 8 show the efficacy of this approach in estimating the perceptual quality of video with complex temporal contents as seen in R_1 and overall performance. It also shows the competitive performance of this algorithm for videos with low temporal complexity, as seen in R_2 and R_3.

3.2.3.2 Channel-induced Distortion

In terms of SROCC, our metrics ranked first on the mobile LIVE database. It also ranked seventh on the CSIQ and fourth on the IVP databases, respectively, with a marginal depreciation (0.15 and 0.13, respectively) from the best performing metrics. On the other hand, in terms of PLOCC, our metric ranked first on both Mobile Live and IVP. It also ranked sixth on the CSIQ database with a depreciation of 0.12 from

the best performing metric. This shows that PeQASO is very suitable for online perceptual evaluation of a wide variety of sequences spanning different spatial and temporal features.

The correlation coefficients of cross-database test are shown in Table 7. The results again show that PeQASO provides a good prediction of the perceptual quality of video with complex temporal contents as seen in R_1. It also shows the competitive performance of this algorithm for videos with low temporal complexity, as seen in R_2 and R_3.

3.2.4 Computational Complexity

In this section, a comparison of the computational complexity of the proposed metric in this work and other metrics is shown. The numbers in the table show average computation times for 120 frames. Those computations were obtained using the same Windows 10 64-bit system with Core(TM) i7-6700K CPU @ 4.00GHz processor, 32.0 GB memory and MATLAB R2015b. We note here that computation times of MATLAB functions involving multidimentional matrices computations vary drastically depending on the the operating systems. Multidimentional matrices calculations have been reported to be significantly faster on Ubuntu systems compared to Windows and macOS systems given the same hardware specifications.

3.2.5 Limitations

The proposed approach relies on the estimation of optical flow maps for the observed frames. Hence, any issue in the optical flow estimation will affect the quality estimator. We noticed a difficulty in estimating the optical flow in the `laser` in the IVP database. This sequence shows a sparkling instruments in a dark background. It starts with an almost still scene with very minor motion caused by a laser beam. Hence, it was difficult for the used Horn-Schunck optical flow estimation algorithm to calculate an accurate optical flow map. However, other optical flow algorithms may

overcome this limitation. This video was excluded from this study. Furthermore, three data points were excluded from all the distorted videos in the three databases in the channel-induced distortion case. This was due to precision issues in the calculations that yielded numerical inconsistencies in the optical flow estimation of the distorted sequences.

Table 9: Computational complexity comparison of the proposed algorithms in this work and other metrics. The numbers in the table show average computation times for 120 frames.

Metirc	Full-Reference						Reduced-Reference		No-Reference	
	PSNR [105]	MS-SSIM [93]	VIF [94]	VSNR [127]	NQM [126]	POTUS (Proposed - Ch. 4)	PeQASO (Proposed)	ST-RRED [113]	Video BLIINDS [114]	VIIDEO [115]
Execution Time (sec)	10.1553	18.5149	255.7289	17.0801	59.4904	15.0296	71.1631	72.4421	422.039	32.3765

CHAPTER IV

PERCEPTUAL QUALITY ASSESSMENT VIA POWER SPECTRAL ANALYSIS

As it was explained in Section 2.3, the HVS has masking characteristics that varies our abilities to perceive certain contents (target) in the presence of other contents (mask). In this section, we take advantage of this phenomenon to model the distortions (target) induced to original content (mask). We built this model by examining the frequency domain components, particularly the power spectral density (PSD). We argue that distortions cause a disruption of the original power spectra which impact users' viewing experiences negatively. The proposed model in this chapter introduces a new theory to model the HVS's sensitivity to distortions by measuring the resulting disruption of the power spectra due to distortions.

One of the main characteristics that define a signal features in the frequency domain is the power spectral density (PSD). In the context of image processing, there are some known correlations between spatial domain characteristics and their impact on the frequency response and power spectrum. For instance, it has been established that smooth spatial regions correspond to low frequency bands, while highly textured areas correspond to the high frequency bands in the power spectrum [58, 59]. Nonetheless, the correspondence and relationship of spatial and temporal video features to the power spectrum is still not very well established in the research community. The works in this chapter is a unique endeavor to investigate this characteristic in the context of perceptual video quality assessment. To the best of our knowledge, this work is the first attempt to explore video quality assessment using power spectral analysis.

The contributions in this chapter are presented as follows. We begin this discussion by introducing a no-reference low-complexity metrics for estimating frame-level distortions in online streaming applications in Section 4.1. Following this work, we discuss in Section 4.2 the details our proposed ubiquitous full-reference framework for estimating sequence-level perceptual video quality via power of tempospatially unified spectral density (POTUS) [141–143].

4.1 No-Reference Frame-Level Distortion Estimation

In this section, we explain our proposed no-reference video quality assessment metric. The proposed approach relies on the fact that any channel-induced distortion will result in a temporal inconsistency between frames within a GOP. We measure this inconsistency through the temporal variation of the PSD across frames. Let f_k and f_{k-1} be the frame of interest and previous frame, respectively. Furthermore, let $\mathcal{S}_k$ and $\mathcal{S}_{k-1}$ denote their respective PSDs:

$$\mathcal{S}_k\left[h,k\right]=\frac{1}{MN}\left|\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}f_k\left[m,n\right]e^{-j2\pi(hm+kn)}\right|^2 \tag{14}$$

where k is the temporal index of the frame in the received video, $M\times N$ is the resolution of the video, and v and u are the discrete frequencies. We next divide the PSD, $\mathcal{S}_k$, into non-overlapping blocks of size $L\times L$. We refer to the PSD of block i in frame f_k as $B_k(i)$. Similarly, $B_{k-1}(i)$ is the PSD of block i in frame f_{k-1}. For every block, we estimate the channel-induced distortion by measuring the energy difference in the temporal domain as follows:

$$\Delta B_k(i)=B_k(i)-B_{k-1}(i). \tag{15a}$$

We next measure the variation of the energy differences within block i in frame f_k as follows:

$$G_k(i)=\frac{\max\left[\Delta B_k(i)\right]}{\sqrt{\mathrm{Var}\left[\Delta B_k(i)\right]}} \tag{15b}$$

where $\max[\cdot]$ is the maximum value in block $\Delta B_k(i)$, $\mathrm{Var}[\cdot]$ is the variance of the values in block $\Delta B_k(i)$, and $G_k(i)$ is the ratio of the maximum PSD value in block i to the standard deviation of the PSD of the block. Next, we compute the negative mean of $G_k(i)$, denoted by $\mathcal{D}_k$, taken over all the spatial indices i in frame k as follows:

$$\mathcal{D}_k = -\mathrm{E}[G_k(i)] \tag{15c}$$

where $\mathrm{E}[\cdot]$ is the expectation operation taken over the spatial indices, i's, for all the blocks. It should be noted that while $B_k(i)$ and $\Delta B_k(i)$ are square matrices, $\mathcal{D}_k(i)$ and $\mathcal{D}_k$ are scalars. Furthermore, the obtained vector for the whole sequence of $\mathcal{D}_k$ values is normalized to obtained $\tilde{\mathcal{D}}_k$. Finally, we amplify the the estimated distortion as follows:

$$\hat{\mathcal{D}}_k = \tilde{\mathcal{D}}_k \cdot \sigma_s(k) \tag{15d}$$

where $\sigma_s(k)$ is the standard deviation of the vector $\left[\tilde{\mathcal{D}}_{k-s}, \cdots \ \tilde{\mathcal{D}}_k, \cdots, \tilde{\mathcal{D}}_{k+s}\right]$. s is the window size, which is determined empirically.

The goal of the operation in (15d) is to scale the measured distortion in (15c) within the context of its neighbouring frames. If the variance of the measured quantity in (15c) is high, this indicates high variations in the PSD levels from one frame to another, which indicates higher error likelihood within the GOP. In our experiments, $s = 5$ and the block size is $L \times L = 16 \times 16$ pixels.

Let us consider a scenario where a frame, k, has been lost and replaced by its predecessor in display order. For this particular frame, (15c) produces $\mathcal{D}_k = 0$. Since $-\infty < \mathcal{D}_k \leq 0$, the normalized value will have values $0 \leq \tilde{\mathcal{D}}_k \leq 1$.

4.1.1 Experiments and Results

All the experiments and tests follow the recommendations published by JCT-VC for common test conditions for HEVC [19]. We use a subset of six difference video sequences in our experiments. All the video sequences were coded using the HEVC

standard using the test model version (HM 12.0) [20]. The coding was done using the main random access profile. Next we detail the coding parameters and the obtained results.

4.1.1.1 Coding Conditions and Simulations Parameters

Table 10 summarizes the sequences used in our experiments and the encoding parameters. We fix the initial Quantization Parameters (QPs) value to 32. For the error patterns, we use the the loss patterns in the proposed NAL unit loss software [144]. The results shown in this part are performed with the 10% loss pattern, which results in 5%-7% loss rate in the tested sequences. In our experiments, only inter-coded frames are subject to losses. Furthermore, Figure 13 shows the spatial information (SI) and temporal information (TI) indices on the luminance channel for the selected sequences, as per the recommendation in [14]. The higher the score on the SI or the TI scale, the more complex the spatial and temporal features of the test sequence. In this context, we diversify the selection of sequences to validate our model under different temporal and spatial features.

Table 10: Test Video Sequences

Sequence	Resolution	Intra Period	FPS	Number of Frames
`RaceHorses`	832x480	24	30	300
`BasketballDrill`	832x480	48	50	500
`PartyScene`	832x480	48	50	500
`BQMall`	832x480	64	60	600
`BasketballDrive`	1920x1080	48	50	500
`ParkScene`	1920x1080	24	24	240

4.1.1.2 Results and Analysis

Figs. 17 and 18 show the calculated measures for `RaceHorses` and `PartyScene` sequences, respectively. From the two plots, we notice that the value of $\hat{D}_k$ peaks at the location of lowest SSIM score. These points correspond to the lost frames, which were

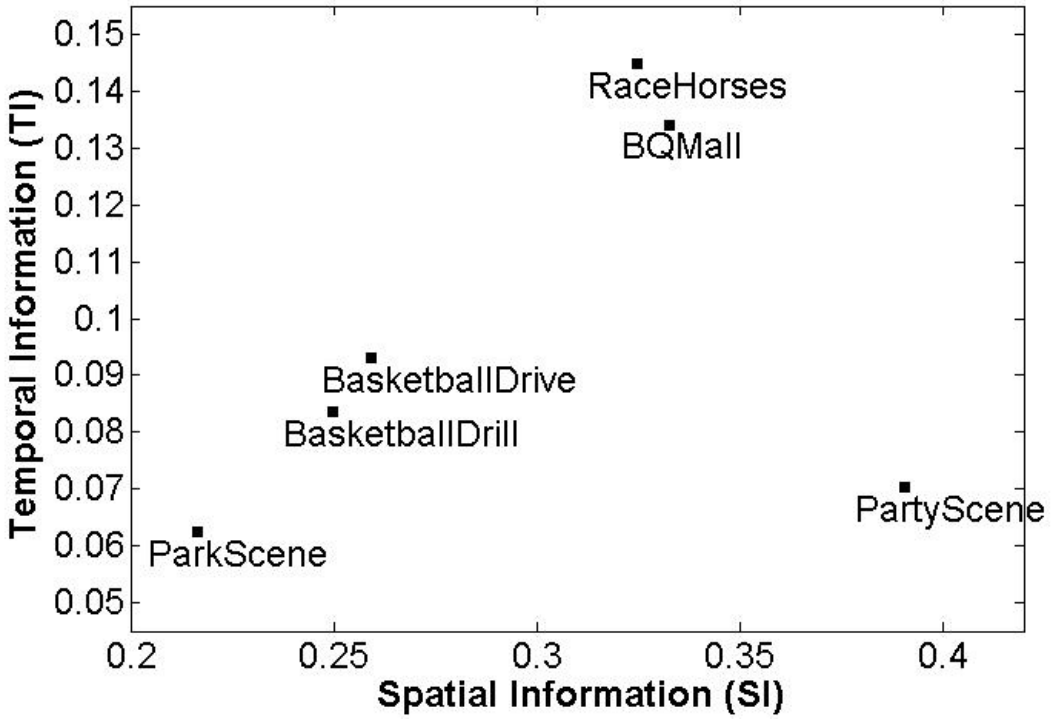


Figure 16: Spatial information (SI) versus temporal information (TI) indices for the selected sequences [14].

replaced by previous frames during the concealment process. In this case, $\mathcal{D}_k \approx 0$, as alluded in Section 3.1.1. This value decreases for the following dependent frames since only a subset of the CTUs in these frames depend on the lost frames.

Table 11: Correlation between the estimated frame distortion, $\mathcal{D}_k$, and the full-reference SSIM values.

Sequences	Correlation Coefficients
RaceHorses	0.79
BasketballDrill	0.76
PartyScene	0.77
BQMall	0.70
BasketballDrive	0.80
ParkScene	0.77

In order to validate the proposed distortion model, we calculate the correlation coefficients between the estimated distortion and the measured SSIM of the corrupted

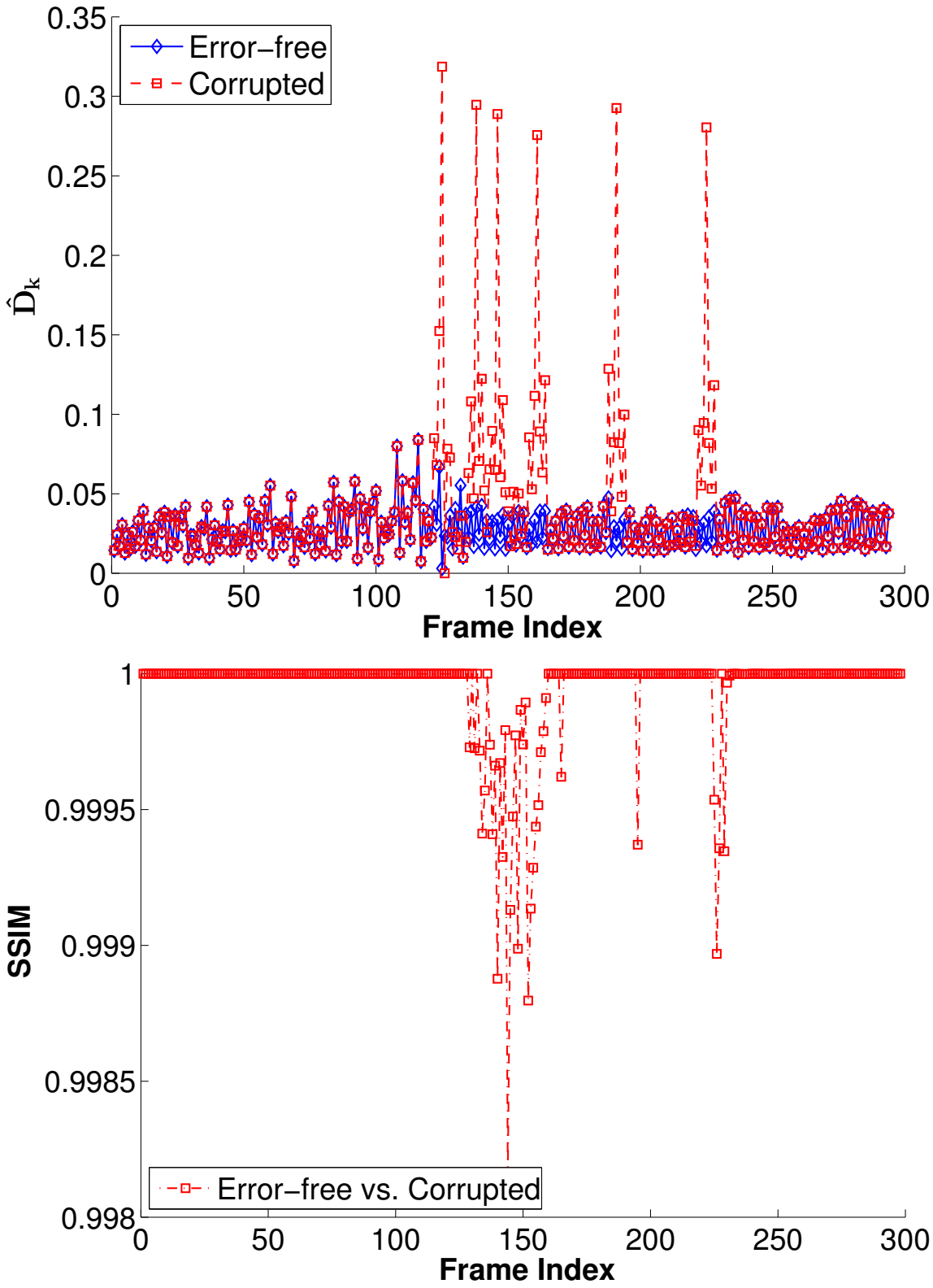


Figure 17: The proposed no-reference quality measure compared with the obtained SSIM for the corrupted and error-free `RaceHorses` sequences.

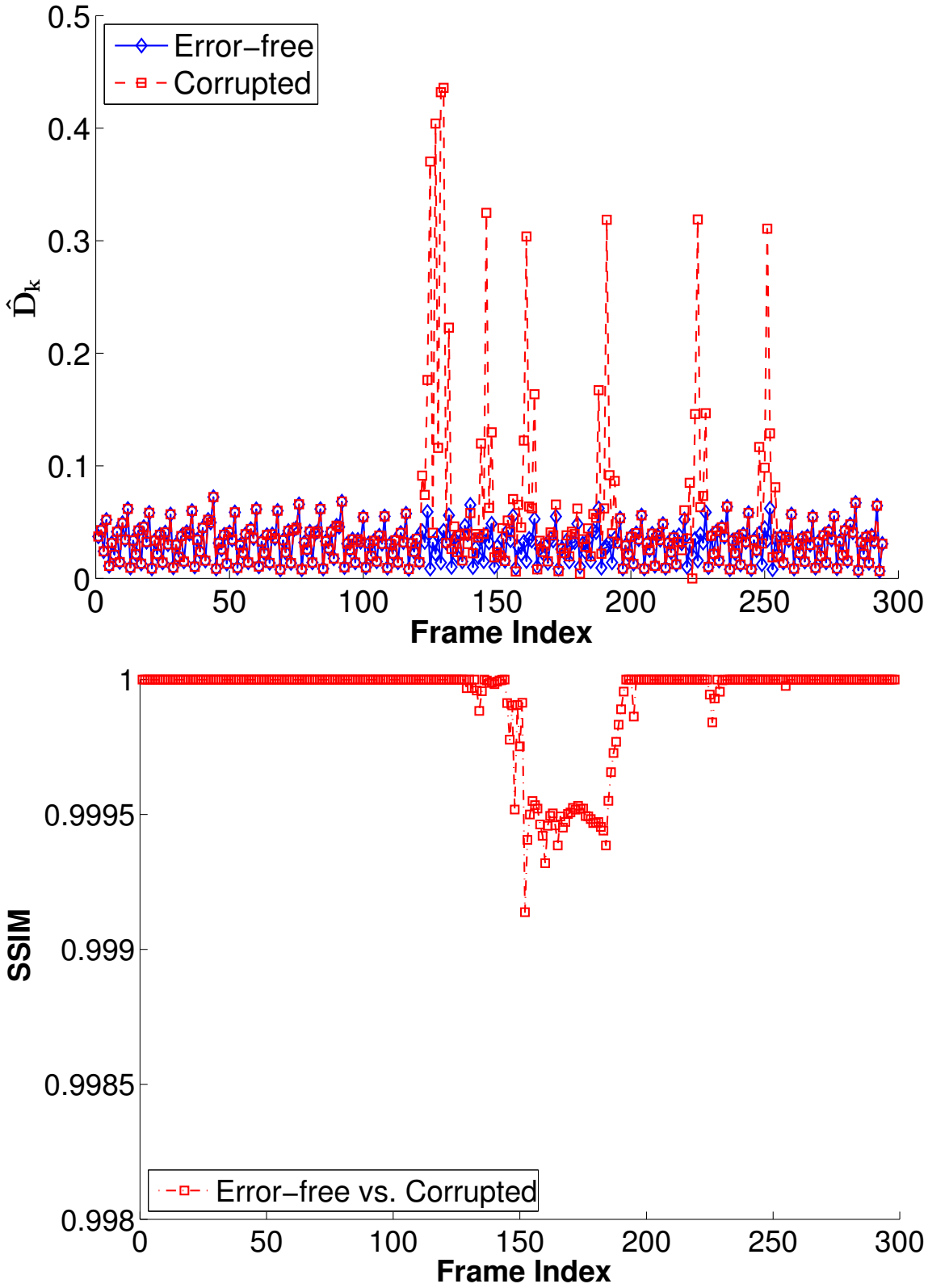


Figure 18: The proposed no-reference quality measure compared with the obtained SSIM for the corrupted and error-free `PartyScene` sequences.

sequence compared with the error-free one. Table 11 summarizes the experimental results for all the tested sequences. Note that the proposed model correlates well with the SSIM values. The correlation coefficients for all test sequences range between 0.70 and 0.80. In particular, the proposed approach works well for the sequences with low temporal complexity such as the `ParkScene` video sequence. In this case, the majority of the changes in the PSDs between consecutive frames is due to the channel-induced distortion. Furthermore, our distortion measure works well for sequences with medium or low temporal complexity, such as `BasketballDrive` and `BasketballDrill`.

The correlation, however, tends to drop for the case of `BQMall` due to the complex nature of localized motion in the video, as can be observed from the TI index in Figure 13. Nonetheless, this problem can be overcome by incorporating spatial inconsistency, which is beyond the scope of this work. The introduced low-complexity metric still performs fairly well for the `RaceHorses` sequence, which is close the `BQMall` in term of spatial and temporal features.

4.2 Power of Tempospatially Unified Spectral Density (POTUS)

4.2.1 3D Power Spectral Density

For any signal, the power spectral density (PSD) describes the distribution of power in the frequency domain for a given signal or time series. Let PSD be defined as $S(\omega)$ as a function of the angular frequency, $\omega = 2\pi f$, measured in radians per second, where f is the frequency in Hz. Hence, the average power over time is given by

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\mathcal{S}(\omega)\,d\omega \tag{16}$$

Let $x(t)$ be a continuous time stochastic process with limited energy, the Fourier

transform is given by

$$X(f) = \mathcal{F}\{x(t)\} = \int_{-\infty}^{\infty} e^{-2\pi i f t} x(t) dt \tag{17}$$

Parseval's theorem establishes the unitary nature of the Fourier Transform. Plancherel theorem then states that the integral of a signal's squared modulus is equal to the integral of the squared modulus of its frequency spectrum. Essentially, this means that the total energy is the aggregate of power over time or spectral power across frequency. This gives a way to estimate the PSD using the signal's Fourier transform. Parseval's theorem is often written as:

$$\int_{-\infty}^{\infty} |x(t)|^2 \, dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |X(\omega)|^2 \, d\omega = \int_{-\infty}^{\infty} |X(2\pi f)|^2 \, df \tag{18}$$

Assuming that the signal $x(t)$ is truncated over the interval $[-T, T]$, the average power over all time is given by the following time average:

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} |x(t)|^2 \, dt = \frac{1}{2\pi} \lim_{T\to\infty} \frac{1}{2T} \int_{-\infty}^{\infty} |X(\omega)|^2 \, d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathcal{S}(\omega) \, d\omega \tag{19}$$

Should this expression hold for continuous increments in the frequency domain, the two terms on the right side of the expression can be written as follows to express the cumulative power:

$$\frac{1}{2\pi} \int_{-\infty}^{\omega} \mathcal{S}(\omega) \, d\omega = \frac{1}{2\pi} \lim_{T\to\infty} \frac{1}{2T} \int_{-\infty}^{\omega} |X(\omega)|^2 \, d\omega. \tag{20}$$

Taking the derivative of the two terms on the right side of this expression yields:

$$\mathcal{S}(\omega) = \lim_{T\to\infty} \frac{1}{2T} |X(\omega)|^2 \tag{21}$$

Let us consider a 3D discrete time-space video signal, $x[m, n, o] \in \mathscr{R}^{M\times N\times O}$, with the one grayscale (luma) channel, where m and n are the spatial indices of the 2D frame and o is the temporal (frame) index. as illustrated in Figure 19. In practice, the frequency response of discrete time sequences is estimated by means of

the Discrete Fourier Transform (DFT). This relationship between $x[m,n,o]$ and its DFT, $X[h,k,l] \in \mathscr{C}^{M\times N\times O}$, is defined as follows:

$$X[h,k,l] = DFT\{x[m,n,o]\} = \frac{1}{\sqrt{MNO}} \sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\sum_{o=0}^{O-1} x[m,n,o]\, e^{-j2\pi\left(\frac{hm}{M}+\frac{kn}{N}=\frac{lo}{O}\right)} \tag{22}$$

$$x[m,n,o] = IDFT\{X[h,k,l]\} = \frac{1}{\sqrt{MNO}} \sum_{h=0}^{M-1}\sum_{k=0}^{N-1}\sum_{l=0}^{O-1} x[m,n,o]\, e^{j2\pi\left(\frac{hm}{M}+\frac{kn}{N}=\frac{lo}{O}\right)} \tag{23}$$

Therefore, Parseval's theorem can be written as follow:

$$\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\sum_{o=0}^{O-1} |x[m,n,o]|^2 = \frac{1}{MNO} \sum_{h=0}^{M-1}\sum_{k=0}^{N-1}\sum_{l=0}^{O-1} |X[h,k,l]|^2 \tag{24}$$

and the PSD, $\mathcal{S}[\phi_h,\phi_k,\phi_l] \in \mathscr{R}^{M\times N\times O}$, is given by

$$\mathcal{S}[\phi_h,\phi_k,\phi_l] = \frac{1}{MNO}|X[h,k,l]|^2, \tag{25}$$

where $\phi_h = 2\pi h/N$, $\phi_k = 2\pi k/M$ and $\phi_l = 2\pi l/O$ are the angular frequency indices in terms of the discrete frequency indices h, k and l.

In order to calculate the average power over time at every spatial frequency, the expression in (25) is integrated over the temporal axis, O. That is

$$\overline{\mathcal{S}}[\phi_h,\phi_k] = \sum_{l=0}^{O} \mathcal{S}[\phi_h,\phi_k,\phi_l], \tag{26}$$

where $\overline{\mathcal{S}}[\phi_h,\phi_k] \in \mathscr{R}^{M\times N}$. Figure 19 illustrates the processing framework for a tensor of frames of size $M \times N \times O$.

$\overline{\mathcal{S}}[\phi_h,\phi_k] \in \mathscr{R}^{M\times N}$ represents a 2D plane containing tempospatial power spectral features of the processed tensor. Given the original and a distorted version, the following section explains the processing and fusion of original and distorted planes to obtain a full-reference perceptual quality metric.

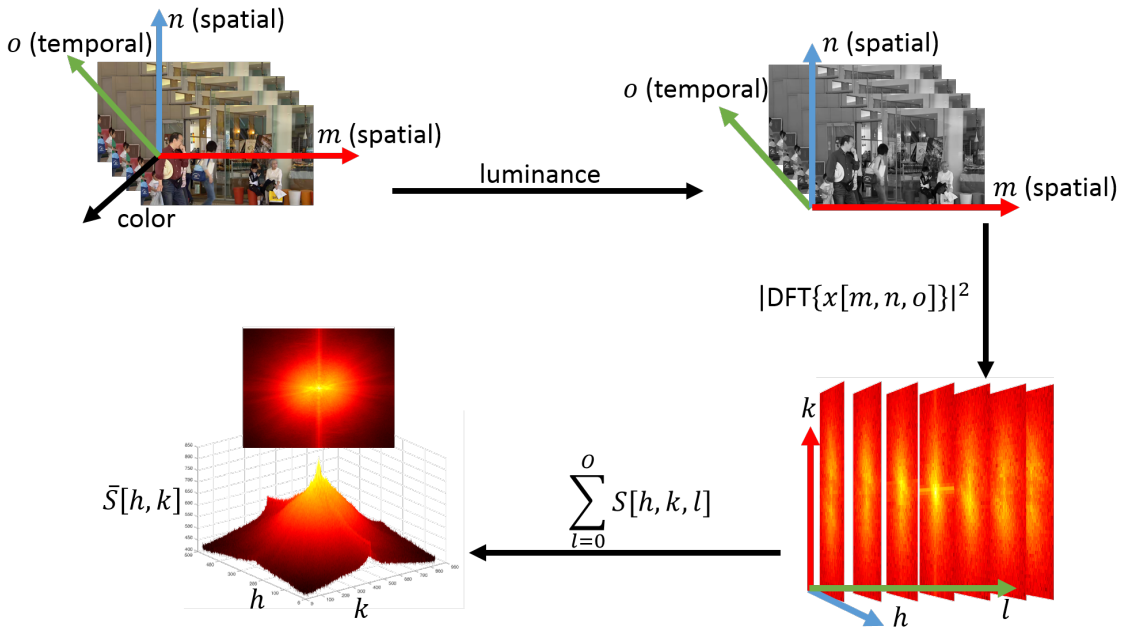


Figure 19: 3D power spectral density tensor-level processing flowchart.

4.2.2 Perceptual Video Quality Assessment via Tempospatial Power Spectral Analysis

In this section, we explain the proposed approach to perceptual video quality assessment via **po**wer of **t**empospatially **u**nified **s**pectral density (**POTUS**). This framework utilizes the 3D tempospatial power spectral densities. Let us consider two videos, an anchor video free of distortion, $x_{\text{ref}}[m,n,o]$, and a distorted version, $x_{\text{rx}}[m,n,o]$. Furthermore, let the video be temporally segmented into a set of equal-size tensors, $M \times N \times O$, where all tensors are composed of the same number of frames. Figure 20 shows a comparison of the tensor-level 2D time-average power spectral density planes for reference and distorted videos. Furthermore, Figure 21 shows a comparison of the tensor-level 2D time-average power spectral density planes for different video contents with varying spatial and temporal features.

Figure 22 shows the incremental change in the tensor-level 2D time-average power spectral density planes for four increasing levels of distortion.

For any given video, let the total number of tensors be T which depends on the

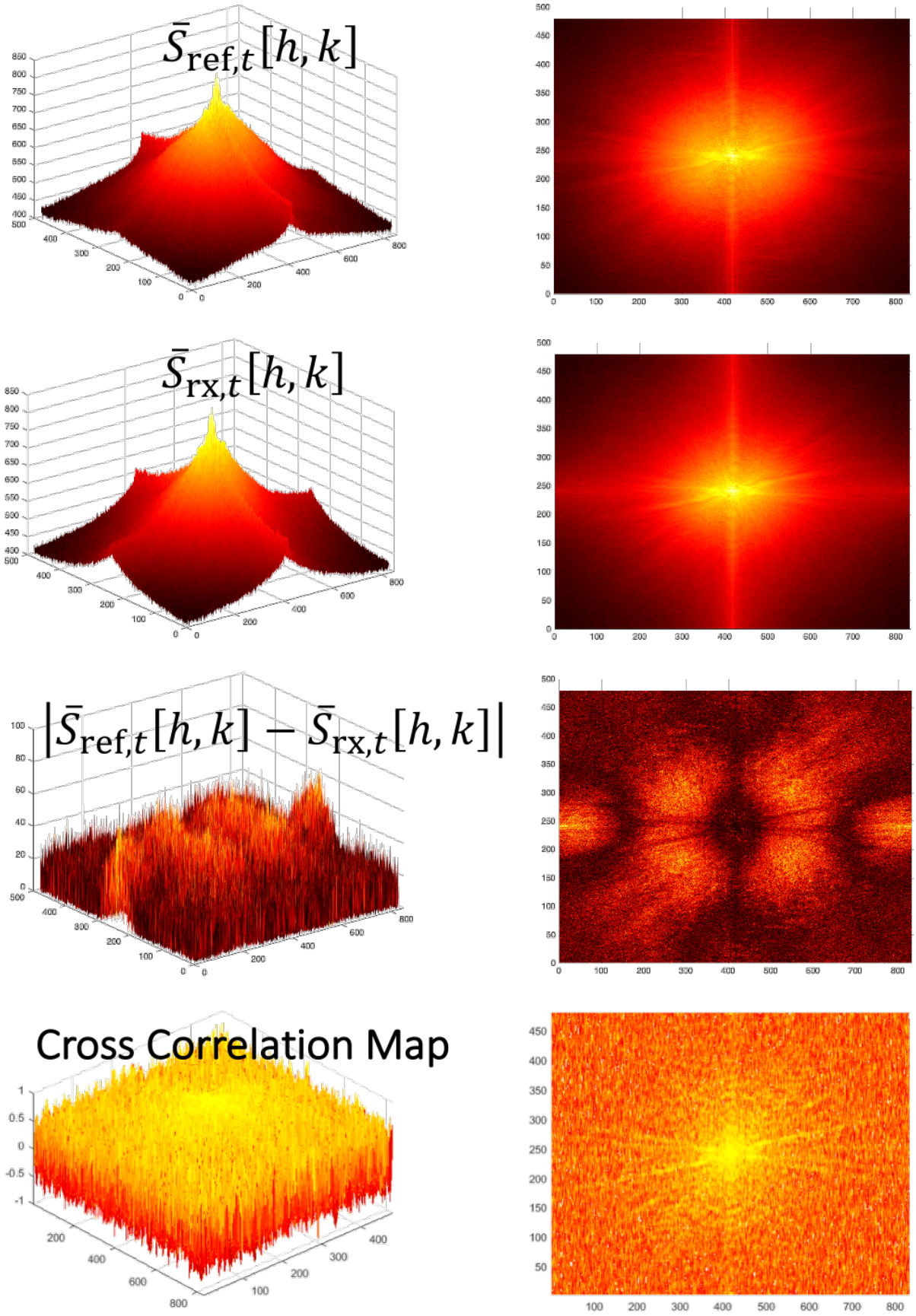


Figure 20: Comparison of the tensor-level 2D time-average power spectral density planes for reference and distorted videos.75

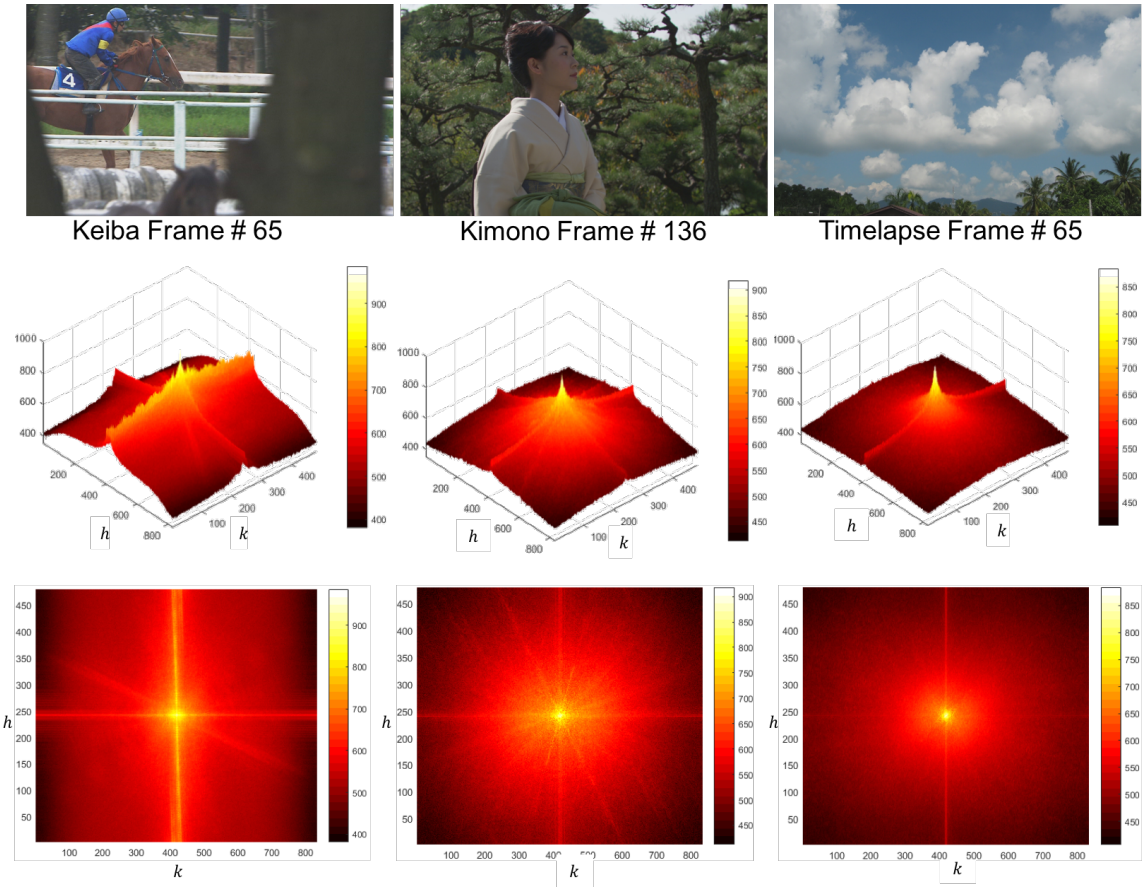


Figure 21: Comparison of the tensor-level 2D time-average power spectral density planes for different video scenes. The center frame of 30-frame tensor is demonstrated on the first row. All three videos were taken from the CSIQ VQA database.

total frame count and tensor size. In this context, we define the cross-correlation between average power spectra maps from the anchor and distorted videos for any tensor $t = 0, \cdots, T-1$, $\zeta_t[\phi_h, \phi_k]$. Every point in the 2D cross-correlation map, $\zeta_t[\phi_h, \phi_k]$, is obtained calculating the *local* cross-correlation of an 11×11 window of neighboring pixels, $\overline{\mathcal{S}}_t \forall \phi_h \in [\phi_h - 5, \phi_h + 5]$ and $\phi_k \in [\phi_k - 5, \phi_k + 5]$. To simplify the notation, we denote this window of neighboring pixels at $[\phi_h, \phi_k]$ of tensor t as $\mathcal{N}_{\phi_h, \phi_k, t}$. Thus, the expression for $\zeta_t[\phi_h, \phi_k]$ is given by:

$$\zeta_t[h,k] = \frac{\sigma_{\overline{\mathcal{S}}_{\text{ref},t} \cdot \overline{\mathcal{S}}_{\text{rx},t}}[h,k] + C}{\sigma_{\overline{\mathcal{S}}_{\text{ref},t}}[h,k] \cdot \sigma_{\overline{\mathcal{S}}_{\text{rx},t}}[h,k] + C} \tag{27}$$

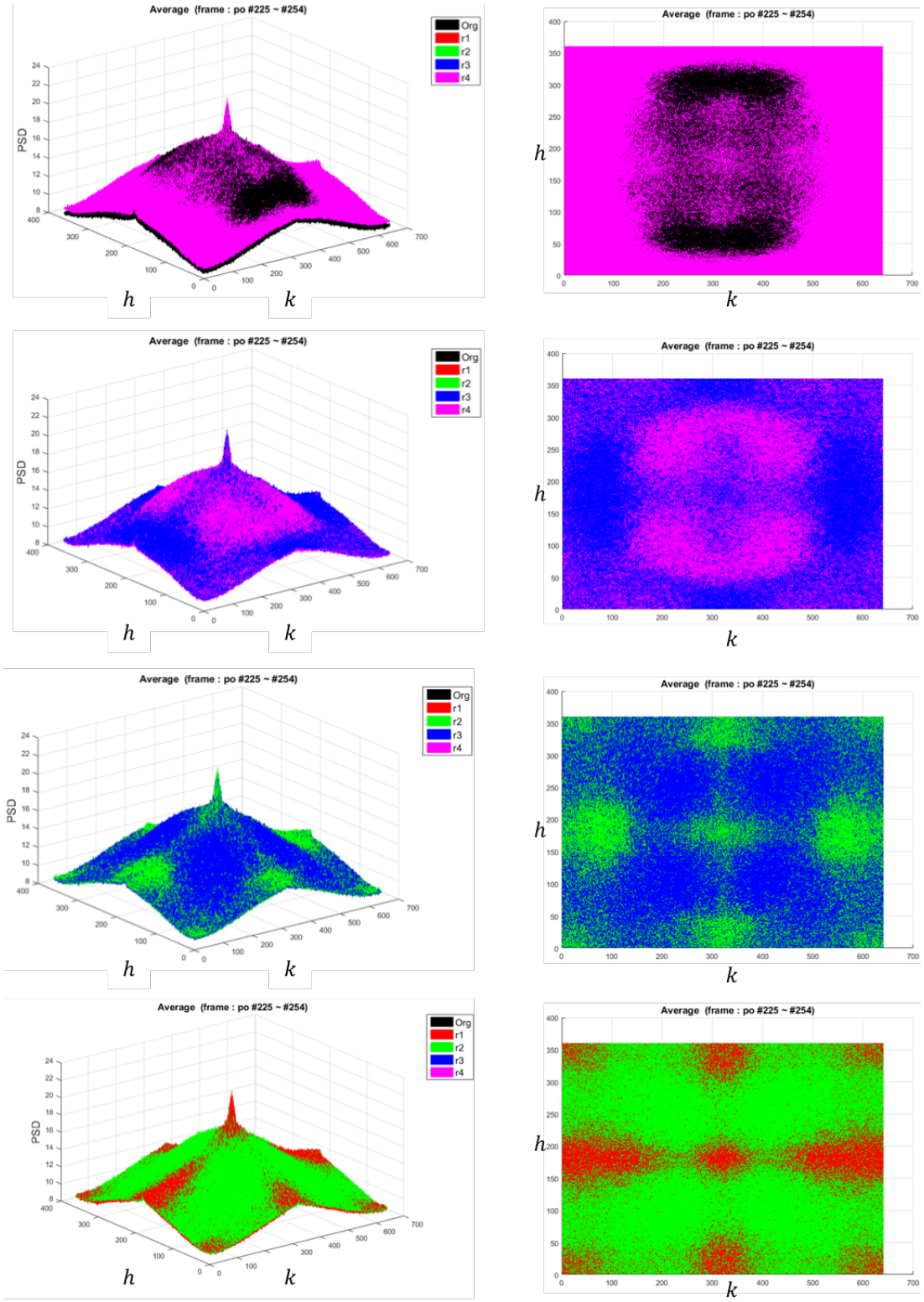


Figure 22: The incremental change in PSD for the same video and same set of frames subject to different distortion levels. This example was taken from the Mobile LIVE database, sequence `Panning Under Oak`, frames 225−254. The distortion magnitudes in the videos are as follows: `r1`>`r2`>`r3`>`r4`>`Org` where `Org` is the anchor video free of distortion.

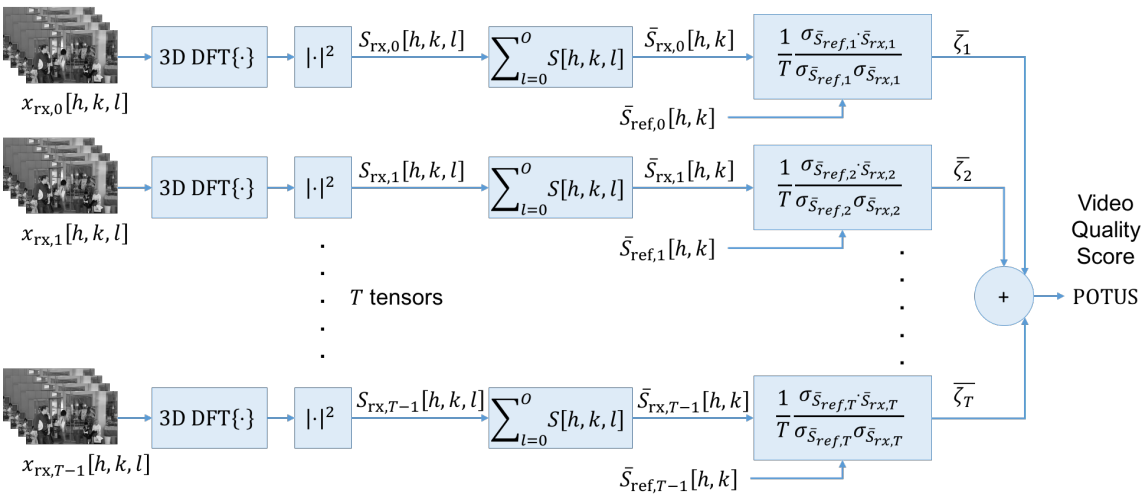


Figure 23: Processing flow chart for the proposed 3D PSD-based perceptual quality metric.

where

$$\sigma_{\overline{\mathcal{S}}_{\mathrm{X},t}}[h,k] = \sqrt{\sum_{u=-d}^{d}\sum_{v=-d}^{d}\omega_{u,v}(\overline{\mathcal{S}}_{\mathrm{X},t}[h+u,k+v] - \mu_{\overline{\mathcal{S}}_{\mathrm{X},t}}[h,k])^2}, \tag{28}$$

$$\sigma_{\overline{\mathcal{S}}_{\mathrm{X},t}\cdot\overline{\mathcal{S}}_{\mathrm{Y},t}}[h,k] = \sum_{u=-d}^{d}\sum_{v=-d}^{d}\omega_{u,v}(\overline{\mathcal{S}}_{\mathrm{X},t}[h+u,k+v] - \mu_{\overline{\mathcal{S}}_{\mathrm{X},t}}[h,k]) \times (\overline{\mathcal{S}}_{\mathrm{Y},t}[h+u,k+v] - \mu_{\overline{\mathcal{S}}_{\mathrm{Y},t}}[h,k]), \tag{29}$$

and

$$\mu_{\overline{\mathcal{S}}_{\mathrm{X},t}}[h,k] = \sum_{u=-d}^{d}\sum_{v=-d}^{d}\omega_{u,v}\overline{\mathcal{S}}_{\mathrm{X},t}[h+u,k+v]\,. \tag{30}$$

where $\sigma_{\overline{\mathcal{S}}_{\mathrm{ref},t}\cdot\overline{\mathcal{S}}_{\mathrm{rx},t}}$ is the cross-covariance, $\mu_{\overline{\mathcal{S}}_{\mathrm{ref},t}}$ and $\mu_{\overline{\mathcal{S}}_{\mathrm{rx},t}}$ are the means, $\sigma_{\overline{\mathcal{S}}_{\mathrm{ref},t}}$ and $\sigma_{\overline{\mathcal{S}}_{\mathrm{rx},t}}$ are the standard deviations of $\overline{\mathcal{S}}_{\mathrm{ref},t}$ and $\overline{\mathcal{S}}_{\mathrm{rx},t}$, respectively. The term ζ_t in (25) defines the tempospatial full-reference perceptual quality for tensor t in the distorted video in terms of its PSD. In our implementation, $C = 4.5 \times 10^{-4}$ is set to prevent instability when denominator is very close zero. In addition, ω is derived from 2D circular symmetric Gaussian weighting function with the window size of 11×11 ($d = 5$). Furthermore,

$$\overline{\zeta_t} = \frac{1}{MN}\sum_{\forall h}\sum_{\forall k}\zeta_t[h,k]\,. \tag{31}$$

Henceforth, the overall video quality, $\mathcal{P}$, is given by the average temporal quality of its tensors. That is

$$\mathcal{P} = \left(\underset{\forall t}{\mathrm{E}} \left[\overline{\zeta_t} \right] \right)^{\beta}, \tag{32}$$

where β is an empirically determined sequence-dependent parameter. Figure 23 shows a complete processing flow chart for the proposed 3D PSD-based perceptual quality metric.

4.2.3 Visual Perception in POTUS

$\zeta_t [h, k]$ is a local cross-correlations map which does not evaluate fidelity, it rather examines the contents in a certain frequency and quantifies the cross-correlation or consistency of contents in that frequency neighborhood with original contents. All the temporal and spatial contents corresponding to a certain frequency are unified within this 2D map. Every frequency spectrum in the original contents emits a certain optical energy to stimulate the HVS. A visual distortion will alter this energy in a certain way depending on the nature and severity of the distortion. This in turn causes discomfort and annoyance to viewers. In the context of visual masking, this framework models the visual sensitivity to distortions by estimating the power spectral cross-correlation, where at every frequency this local cross-correlation estimates the human visual discomfort in that frequency neighborhood. Quantifying the cross-correlation of spectral data in every frequency neighborhood measures the masking effect of the original contents (mask) in the presence of distortion (target). In other words, the local cross-correlation acts as a measure of annoyance or discomfort due to disruption of the original power spectrum caused by induced distortion. A high positive correlation indicate the contents to be similar which yields little to no distortion to the viewer. Low positive and negative local cross-correlation values indicate a degradation in perceptual quality due to distortion. By averaging the map to obtain $\overline{\zeta_t}$, we incorporate the contribution to discomfort from every frequency.

This averaging operation penalizes frequency spectra with low positive and negative cross-correlations by reducing the overall average for the whole tensor's perceptual quality.

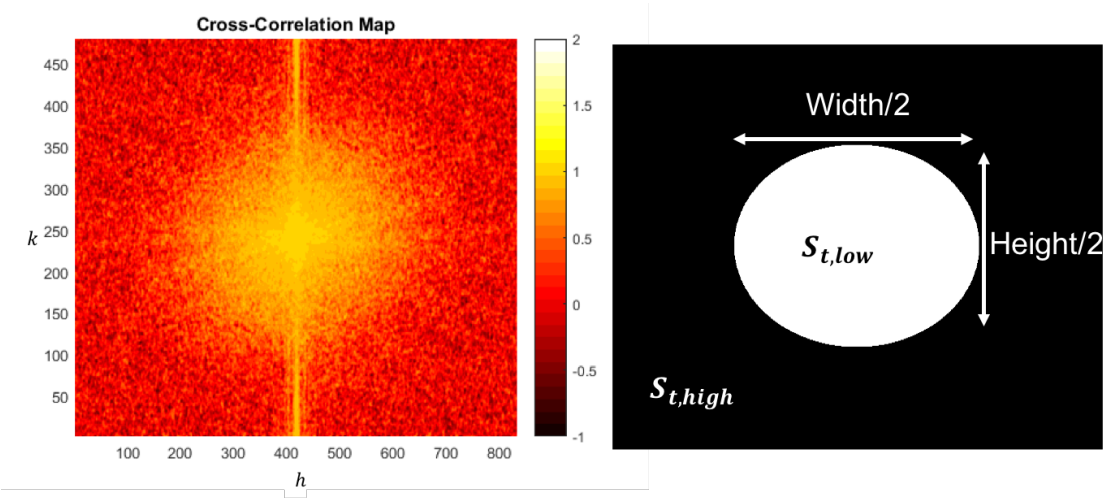


Figure 24: Normalization filter for numerical calibration of various spectrum ranges.

4.2.3.1 Normalization for Spectra Variability

Table 12: Normalizations parameters values for the three databases.

Database	α	β	γ
LIVE Mobile	1	5	0.4
CSIQ	1.9	1	0.9
IVP	1	3	0.59

The numerical range and characteristics of the PSD depends on the tempospatial features in the scene. The variations of temporal and spatial features and contents in different videos produce different PSD features with possibly different numerical ranges. To account for this variability in numerical ranges of the PSD of different tensors, a normalization and calibration step is required to ensure the accuracy of this metric. This step is performed a follows.

First, the mean of low frequency components, $\overline{\zeta}_{t,\text{low}}$, and mean low frequency bands, $\overline{\zeta}_{t,\text{high}}$, are calculated. Figure 24 shows the filter applied to $\zeta_t\,[h,k]$ to separate

the two regions in the frequency domain. The two means are used to calculate the following ratios:

$$\overline{\zeta}_{t,\text{ratio}} = \frac{\overline{\zeta}_{t,\text{low}}}{\overline{\zeta}_{t,\text{high}}}, \tag{33a}$$

$$\overline{\zeta}_{t,\text{norm ratio}} = \frac{\overline{\zeta}_{t,\text{ratio}}}{\max\limits_{\forall \text{ distortions}} \left\{\overline{\zeta}_{t,\text{ratio}}\right\}} \tag{33b}$$

where $\max\limits_{\forall \text{ distortions}} \left\{\overline{\zeta}_{t,\text{ratio}}\right\}$ is obtained from the set of all distorted tenors obtained from distorted versions of a given video. The tensor level normalized quality score is then given by:

$$\eta_t = \begin{cases} \overline{\zeta}_t - \alpha \left(\overline{\zeta}_{t,\text{ratio}}\right)^{\beta}, & \overline{\zeta}_{t,\text{norm ratio}} > \gamma \\ \overline{\zeta}_t, & \text{otherwise} \end{cases}. \tag{33c}$$

Table 12 shows the values of the α, β, and γ parameters used for every database. Finally, the expression in (32) is updated to obtain the overall video quality, which is given by the average temporal quality of its tensors:

$$\mathcal{P} = \mathrm{E}\left[\eta_1, \eta_2, \cdots, \eta_T\right]. \tag{33d}$$

4.2.4 Experiments and Results

In this section, we discuss the experimental results and validation for the proposed perceptual quality estimation framework. We tested the proposed framework on all the distortion types included in three independent VQA databases [15, 71, 79] detailed in 3.2.1. We show the results of each of these databases independently and the overall accuracy of the perceptual quality estimation across databases when applicable.

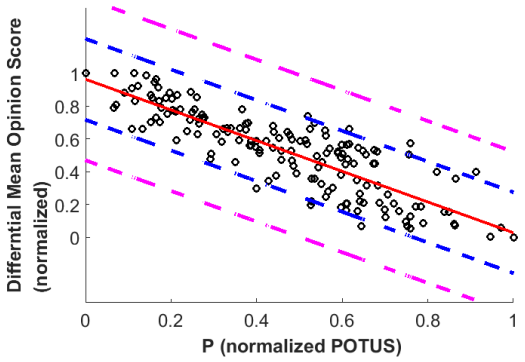
Figure 25 show the scatter plots of POTUS versus DMOS scores in the three databases. Furthermore, Tables 13-14 show the correlation coefficient for the POTUS in terms of SROCC and PLOCC for the three tested databases. In general, POTUS performed very well and was competitive with all other compared metrics. The correlation coefficients in Tables 13-14 show that the POTUS leads the scores in

Table 13: SROCC for the proposed metric and a set of the most common video and images quality assessment metrics tested on all sequences in the three databases.

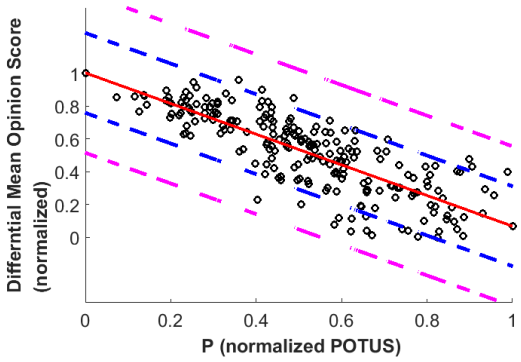
Database	Distortion	Spearman Correlation Coefficient									
		Full Reference									
		PSNR [105]	VQM [107]	MOVIE [72]	TQV [129]	ViS3 [15]	MS-SSIM [93]	VIF [94]	VSNR [127]	NQM [126]	POTUS (Proposed)
LIVE Mobile	Compression	0.819	0.772	0.774	0.764	0.757	0.804	0.861	0.874	0.850	**0.959**
	Wireless	0.793	0.776	0.651	0.754	0.845	0.813	0.874	0.856	0.899	**0.952**
	Rate adaptation	0.598	0.648	0.720	NA	NA	0.738	0.639	0.674	0.678	**0.879**
	Temporal dynamics	0.372	0.386	0.158	NA	NA	0.397	0.124	0.317	0.238	**0.811**
	All	0.678	0.695	0.642	NA	NA	0.743	0.744	0.752	0.749	**0.858**
CSIQ	H.264	0.802	0.919	0.897	0.955	0.920	0.679	**0.984**	0.637	0.773	0.922
	WLPL	0.851	0.801	0.886	0.842	0.856	0.757	**0.954**	0.889	0.866	0.872
	MJPEG	0.509	0.647	0.887	0.870	0.789	**0.900**	0.895	0.403	0.038	0.89
	SNOW	0.759	0.874	0.900	0.831	0.908	0.854	0.885	0.836	0.850	**0.91**
	AWGN	0.906	0.884	0.843	0.908	**0.928**	0.922	0.917	0.896	0.759	0.911
	HEVC	0.785	0.906	0.933	0.902	0.917	0.939	**0.944**	0.852	0.874	0.906
	All	0.579	0.789	0.806	0.814	**0.841**	0.752	0.582	0.635	0.523	0.835
IVP	DIRAC	0.860	0.891	0.888	0.786	0.926	0.811	0.677	0.798	0.744	**0.941**
	H.264	0.866	0.862	0.823	0.672	0.876	0.687	**0.903**	0.771	0.733	0.797
	IPPL	0.711	0.650	0.858	0.629	0.807	0.620	0.690	0.766	0.795	**0.912**
	MPEG-2	0.738	0.791	0.823	0.557	0.834	0.654	0.652	0.639	0.584	**0.89**
	All	0.728	0.845	0.880	0.701	**0.896**	0.580	0.305	0.666	0.780	0.79

Table 14: PLOCC for the proposed metric and a set of the most common video and images quality assessment metrics tested on all sequences in the three databases.

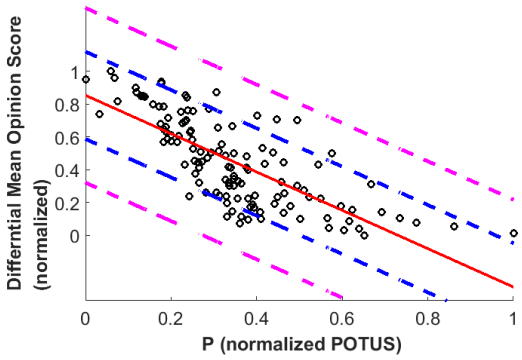
Database	Distortion	Pearson Correlation Coefficient									
		Full Reference									
		PSNR [105]	VQM [107]	MOVIE [72]	TQV [129]	ViS3 [15]	MS-SSIM [93]	VIF [94]	VSNR [127]	NQM [126]	POTUS (Proposed)
LIVE Mobile	Compression	0.784	0.782	0.810	0.788	0.773	0.766	0.883	0.849	0.832	**0.951**
	Wireless	0.762	0.791	0.727	0.754	0.845	0.771	0.898	0.849	0.874	**0.949**
	Rate adaptation	0.536	0.591	0.681	NA	NA	0.709	0.664	0.658	0.677	**0.856**
	Temporal dynamics	0.417	0.407	0.244	NA	NA	0.407	0.105	0.427	0.365	**0.8**
	All	0.691	0.702	0.716	NA	NA	0.708	0.787	0.759	0.762	**0.85**
CSIQ	H.264	0.835	0.916	0.904	0.965	0.918	0.658	**0.989**	0.607	0.779	0.937
	WLPL	0.802	0.806	0.882	0.784	0.850	0.761	**0.963**	0.899	0.889	0.832
	MJPEG	0.460	0.641	0.882	0.871	0.800	0.864	**0.915**	0.390	0.006	0.892
	SNOW	0.769	0.840	0.898	0.846	**0.908**	0.845	0.899	0.859	0.845	0.883
	AWGN	0.949	0.918	0.855	0.930	0.916	0.894	**0.954**	0.921	0.779	0.926
	HEVC	0.805	0.915	0.937	0.913	0.933	0.886	**0.959**	0.848	0.886	0.901
	All	0.565	0.769	0.788	0.795	**0.830**	0.655	0.535	0.610	0.484	0.811
IVP	DIRAC	0.878	0.898	0.870	0.811	**0.936**	0.759	0.748	0.798	0.745	0.888
	H.264	0.855	0.869	0.845	0.744	0.898	0.662	**0.909**	0.735	0.746	0.757
	IPPL	0.673	0.642	0.842	0.735	0.802	0.638	0.633	0.721	0.736	**0.82**
	MPEG-2	0.718	0.836	0.824	0.533	**0.912**	0.712	0.722	0.624	0.599	0.813
	All	0.723	0.847	0.879	0.722	**0.896**	0.548	0.248	0.653	0.737	0.75



(a) Mobile LIVE Database POTUS scatter plot



(b) CSIQ database POTUS scatter plot



(c) IVP Database POTUS scatter plot

Figure 25: Scatter plot of the perceptual quality predictor, $\mathcal{P}$, versus the reported DMOS in the three subjective quality video databases. The blue and pink lines are $P \pm \sigma$ and $P \pm 2\sigma$, respectively, where σ is the data standard deviation. The red line is the mean.

all distortion groups in the LIVE Mobile database in terms of both SROCC and PLOCC. Furthermore, POTUS performed competitively in the CSIQ databases with a marginal depreciation of 0.006 and 0.019 behind the best performing metric in the database for SROCC and PLOCC, respectively. In the IVP database, the proposed metric fell behind the best performing metric by 0.106 and 0.145 in terms of SROCC and PLOCC, respectively. We note herein that these results were obtained without curve fitting or any processing beyond the reported formulation.[1]

[1]Please refer to Table 9 in Section 3.2.4 for the computational complexity comparison.

4.2.4.1 Compression Artifacts Performance

In general, POTUS performed well across different variations of compression artifacts spanning the three tested databases. In terms of SROCC and PLOCC, the average correlations of all distortion classes are 0.902 and 0.878, respectively. Considering only H.264 compression, the average correlation is 0.893 and 0.882 for SROCC and PLOCC, respectively.

4.2.4.2 Channel-Induced Distortions Performance

POTUS also performed well across different variations of channel-induced distortions in the three databases. In terms of SROCC and PLOCC, the average correlations of all distortion classes are 0.912 and 0.882, respectively.

4.2.4.3 3D Spectra Spatial and Temporal Decomposition

The advantage of using 3D PSD in the proposed approach is including both spatial and temporal features towards the perceptual evaluation of video (tensor) quality. In order to further examine the significance of spatial and temporal features in this process, we further analyze our tempospatial tensor by utilizing the proposed 3D spectra decomposition in Long and AlRegib [145]. Therein, the authors propose decomposing a 3D FFT spectrum into two parts: a component related to temporal variations and a component related to spatial variations. This decomposition is based on spectral locations. For any given spectral point $X[h,k,l]$ in (22), the spatial and temporal components are given by projecting the point onto the spatial and temporal axes, respectively. Therefore, the spatial decomposition, $X_s[h,k,l]$, and temporal decomposition $X_s[h,k,l]$, are given by:

$$X_s[h,k,l] = X_t[h,k,l] \times \frac{\sqrt{h^2+k^2}}{\sqrt{h^2+k^2+l^2}} \tag{34a}$$

$$X_t[h,k,l] = X_t[h,k,l] \times \frac{l}{\sqrt{h^2+k^2+l^2}} \tag{34b}$$

We use these two components to examine the performance of the spatial and temporal components of POTUS. on the videos in the Mobile LIVE database. The results in Table 15 show the correlations coefficients with the DMOS scores of the spatial and temporal components. One can observe that while the spatial correlations are always higher, the temporal correlations are very close. This also shows the human visual system's distraction by spatial distortions more than temporal ones.

Table 15: Spatial and temporal decomposition correlation coefficients.

Distortion	Spearman Correlation Coefficient		
	Spatial	Temporal	Uncomposed
Compression	0.826	0.777	0.9586
Wireless	0.837	0.730	0.9518
Rate adaptation	0.591	0.444	0.8789
Temporal dynamics	0.561	0.449	0.8113
All	0.721	0.632	0.8576
Distortion	Pearson Correlation Coefficient		
	Spatial	Temporal	Uncomposed
Compression	0.816	0.752	0.9513
Wireless	0.825	0.735	0.9488
Rate adaptation	0.548	0.383	0.8562
Temporal dynamics	0.550	0.423	0.7988
All	0.706	0.618	0.8501

CHAPTER V

FEATURES TO PERCEPTION: TEMPOSPATIAL POOLING STRATEGIES FOR DIFFERENT VISUAL DISTORTION MAPS

5.1 Tempospatial Video Pooling for PVQA

A raw video is a large data structure considering the time axes and three colour channels. In most applications, we tend to simplify this structure by ignoring the color components (chrominance) and process the luminance channel only. Nonetheless, this single-channel video remains a huge three dimensional data structure that can be processed, segmented and pooled following several strategies. Furthermore, most of the work so far suggest compacting this large data structure into a single floating number to assign a quality score. Hence, there is a need for a concrete understanding of the statistical properties of visual feature maps used in such algorithms. This section will focus on the significance of statistical pooling strategies, at the spatial and the temporal levels, and its correlation with the visual feature maps in video processing.

There are several ways to perform statistical pooling of video features and data. We define the process of pooling video features in both the spatial and temporal (time) domains as tempospatial processing. There are two major approaches to perform tempospatial pooling. First, there is *intra-frame spatial processing* where spatial pooling is performed for every frame independently first to obtain frame-level descriptors. These descriptors are then used to perform temporal pooling. This is the most common approach in video quality assessment because of simplicity and computational efficiency. Secondly, other approaches propose *inter-frame spatial processing*

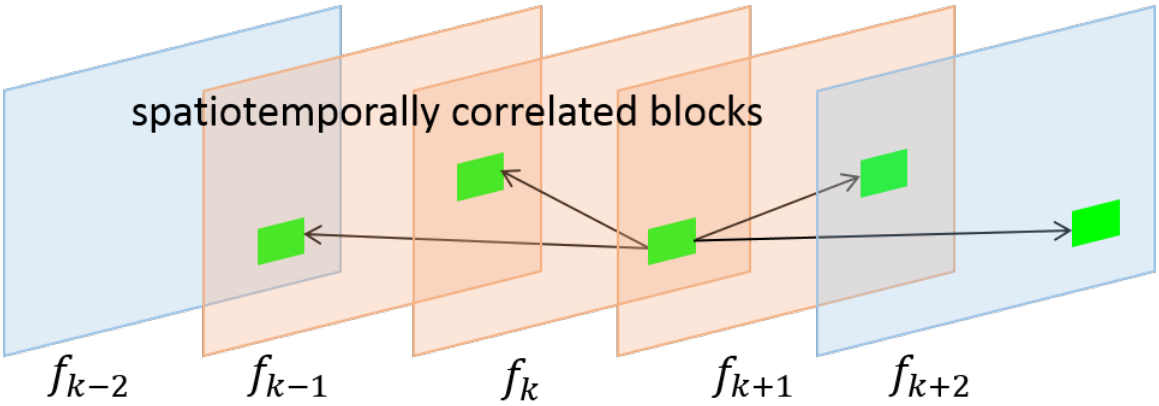


Figure 26: Inter-frame tempospatial processing of temporally correlated blocks.

of correlated blocks where we perform statistical pooling of different combinations of temporally correlated spatial slices (blocks) across frames. The cube, which is a composite of the correlated spatial blocks, is then processed and pooled to produce a single or multiple descriptors.

We focus on the first approach, namely, intra-frame spatial processing. The goals are twofold. First, we show the efficacy of different statistical moments (pooling) for different distortion maps. Secondly, we examine the performance of different distortion maps over different distortion types and databases. We look at the performance of three full-reference pixel-level distortion maps: squared error, local SSIM values for each pixel and absolute difference of optical flow.

Let the frame of interest be f_k with a resolution of $M \times N$. The first distortion map examines the fidelity of pixels. Thus, the first spatial distortion map is given by:

$$\mathbf{D}_k^{\mathrm{SE}}[m,n] = \left[\mathbf{f}_k^{\mathrm{rx}}[m,n] - \mathbf{f}_k^{\mathrm{ref}}[m,n]\right]^2 \qquad (35)$$
$$\forall m \in [0, M-1], n \in [0, N-1],$$

where $\mathbf{f}_k^{\mathrm{ref}}$ and $\mathbf{f}_k^{\mathrm{rx}}$ are the luminance channels of the reference and received frame, respectively. The second spatial distortion map measures the local similarity structure at the pixel level. It is given by [146]:

$$\mathbf{D}_k^{\mathrm{SSIM}}[m,n] = \qquad (36)$$

$$\frac{(2\mu_{\text{ref}}[m,n]\mu_{\text{rx}}[m,n]+c_1)(2\sigma_{(\text{ref,rx})}[m,n]+c_2)}{(\mu^2_{\text{ref}}[m,n]+\mu^2_{\text{rx}}[m,n]+c_1)(\sigma^2_{\text{ref}}[m,n]+\sigma^2_{\text{rx}}[m,n]+c_2)},$$

where μ_{rx}, μ_{ref}, σ_{rx}, σ_{ref}, and $\sigma_{(\text{ref,rx})}$ are the local means, standard deviations, and cross-covariance for images reference frame, ref, and received one, rx. c_1 and c_2 are regularization constants. The third feature map is based on the optical flow and it captures the deviation within the motion field of the received and distorted frame from the original one. Let $\mathbf{U}_k$ and $\mathbf{V}_k$ denote the matrices of the horizontal and vertical optical flow velocities, respectively, of frame, f_k. Furthermore, let $\mathbf{R}_k = \sqrt{\mathbf{U}_k^2+\mathbf{V}_k^2}$ denote the matrix of magnitudes of the flow velocities [135]. The motion field distortion map is given by:

$$\mathbf{D}_k^{\text{OF}}[m,n] = |\mathbf{R}_k^{\text{rx}}[m,n] - \mathbf{R}_k^{\text{ref}}[m,n]| \tag{37}$$

where $\mathbf{R}_k^{\text{rx}}$ and $\mathbf{R}_k^{\text{ref}}$ are the optical flow maps of the received and reference frames, respectively. All the results in this paper were obtained using the Horn-Schunck optical flow method [135].

Figure 27 shows visual examples from `Chipmunks` sequence in the CSIQ Video-Quality database [15] that shows the three distortion maps used in this study.

In the following sections, we show the data of eight spatial pooling operations and nine temporal pooling operations for all three distortion maps. In the spatial domain, we focus on the following statistics: `mean`, `standard deviation`, `mean to standard deviation ratio`, l_1`-norm`, l_2`-norm`, `maximum`, `kurtosis` and `skewness`. In the temporal domain operations, we add the `median`.

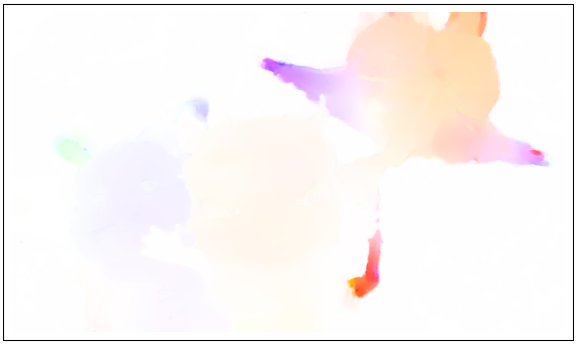
5.2 Experiments and Results

5.2.1 Databases and Test Videos

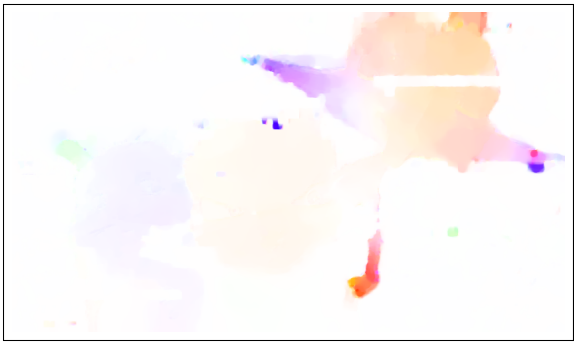
Table 16 shows a summary of the three databases used in this paper and their video contents. Furthermore, Figure 28 shows the temporal information (TI) and spatial

(a) Anchor frame 37

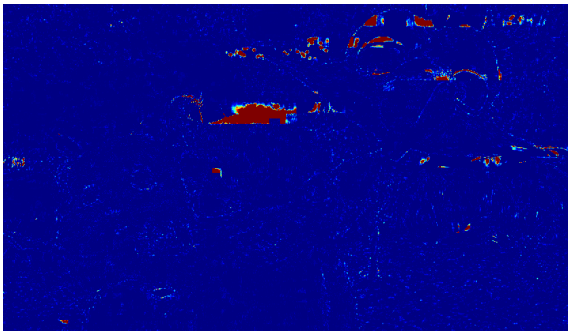
(b) Distorted frame 37



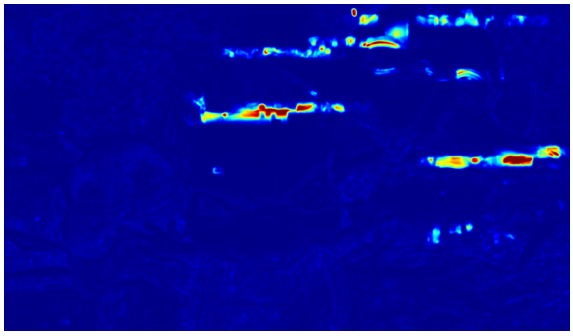
(c) Optical flow of anchor frame



(d) Optical flow of distorted frame



(e) Squared residual map: $\mathbf{D}_k^{\mathrm{SE}}$



(f) SSIM distortion map: $\mathbf{D}_k^{\mathrm{SSIM}}$

(g) Optical flow map residual, $\mathbf{D}_k^{\mathrm{OF}}$

Figure 27: Example from video `Chipmunks` in the CSIQ Video-Quality database [15] that shows the three distortion maps used in this study. The SSIM value of the distorted frame to the anchor is 0.97 and the PSNR value is 31.55 dB.

Table 16: Summary of the VQA databases used in this study and their video contents.

Database	No. of Sequences	No. of Distortions	Total no. of Distorted Videos	Resolution	Duration (s)	Frame Rate (fps)
Mobile LIVE [71]	10	6	197	1280 × 720	10	25
CSIQ [15]	12	6	216	832 × 480	10	24-60
IVP [79]	10	4	128	1920 × 1088	10	25

information (SI) indices for the three evaluated databases. These plots where obtained using the P.910 subjective video quality assessment recommendation of the TU Telecommunication Standardization Sector [14].

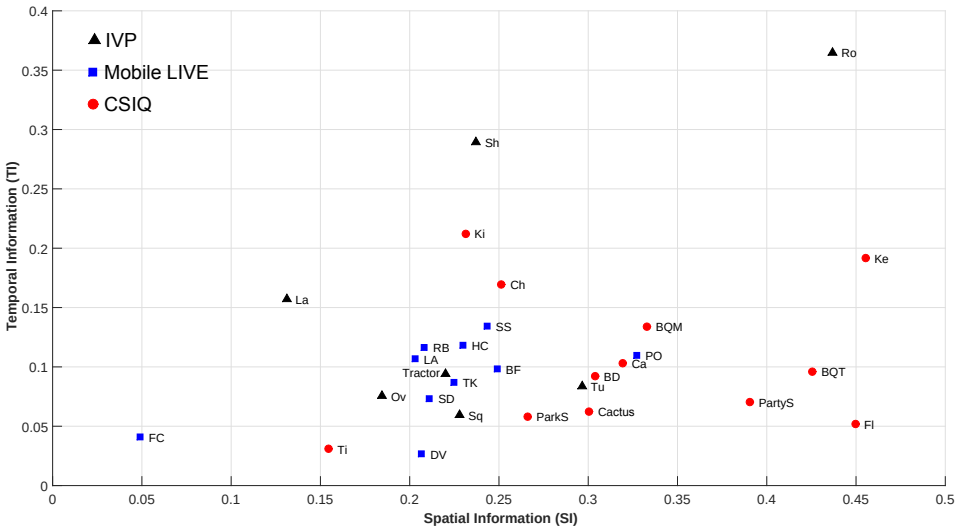


Figure 28: Spatial information (SI) versus temporal information (TI) of the tested sequences from the the three used VQA databases.

5.2.2 Results and Analysis

Tables 19, 23, and 27 in Appendix 5.A show the top three most effective operations, with respect to the Spearman's rank-order correlation coefficient (SROCC), applied on the squared error distortion map, the structure similarity map, and the optical flow map, respectively. Furthermore, Tables 20, 24, and 28 show the top three most effective operations, with respect to the Pearson's linear-order correlation coefficients (PLOCC). The results are shown separately for every class of distortion in three databases [15, 71, 79]. Additionally, Tables 19-28 show the number of videos used

to obtain the correlation scores. In all the cases, we used the maximum number of videos provided in every database. Every cell reports the following: (*i*) the used spatial pooling strategy, (*ii*) the used temporal pooling strategy, (*iii*) the SROCC value, and (*iv*) the average number of frames per sequence.

In order to understand these raw tables and large data, further analysis were performed on Tables 19, 20, 23 24, 27, 28 (Appendix 5.A) to uncover significant patterns in these results. We report a summary of the results and findings in terms of both, SROCC and PLOCC in Tables 17-18.

In temporal pooling, we used all the frames in the video with two exceptions. In the cases of frame-freezes (live feed) and temporal dynamics in the Mobile LIVE [71], we performed the pooling over last 50% and 60% of the frames, respectively. In the frame-freeze case, this was done to eliminate the redundant frames due to buffering. In the temporal dynamics case, this was done to maximize the correlation coefficients as the DMOS values are more biased towards the quality of last few seconds of the video. The entries highlighted with bold text indicate the best performing distortion map among all three features.

5.2.2.1 **Distortion Maps Performance**

Studying Tables 19, 20, 23 24, 27, 28 (Appendix 5.A), we make several key observations.

- For all compression artifacts, the $\mathbf{D}_k^{\text{SSIM}}$ map works best for four out of nine total test sets. The optical flow, $\mathbf{D}_k^{\text{OF}}$, ranks second with three sets.
- For H.264 compression in particular, $\mathbf{D}_k^{\text{SSIM}}$ superseded the other two and performed best when all the sequences from all databases were pooled in one set.
- For the case of channel-induced distortions, $\mathbf{D}_k^{\text{OF}}$ performed best followed by $\mathbf{D}_k^{\text{SE}}$. The $\mathbf{D}_k^{\text{OF}}$ map yielded the best correlation with DMOS values when all the sequences were pooled together.

- The tables show a total of 18 test sets: six from LIVE, six from CSIQ, four from IVP, and two cross-database sets. In general, the optical flow map, $\mathbf{D}_k^{\mathrm{OF}}$, performed best in seven test sets. The SSIM map, $\mathbf{D}_k^{\mathrm{SSIM}}$, came in second performing best in six sets. Finally, $\mathbf{D}_k^{\mathrm{SE}}$ performed best in five test sets.

- In the cross-database tests, $\mathbf{D}_k^{\mathrm{OF}}$ performed best for H.264 channel-induced distortions, and SSIM performed best with H.264 compression.

- By looking at the correlation coefficients values across Tables 19,23, 27, the squared error is considered the weakest if we exclude MJPEG compression. While it performed best for 5 video sets including MJPEG compression, the correlation coefficients of the other four sets using the other two feature maps are very competitive. This is an indication that ***human perception is mostly sensitive to motion and spatial structures in videos.*** Hence, any artifact or distortion affecting motion fields and spatial structures are expected to reduce video perceptual quality.

Furthermore, looking at the results of similar distortions across databases, we noticed a variation in terms of efficacy of feature maps and pooling strategies. We discuss here the case of H.264 compression artifacts with respect to $\mathbf{D}_k^{\mathrm{SE}}$ operations in 19. Spatial domain operations are consistent across all three databases and the cross-database validation. The `max` and `kurtosis` are clearly the three most effective spatial pooling operations in this case. However, the temporal pooling operations do not follow the same behaviour. The LIVE database produced the `max` and `mean`, the CSIQ database produced `skewness` and `kurtosis`, and finally the IVP database produced `mean` and l_2`-norm`. The cross-database results show that l_2`-norm` and `mean` are the most effective operations when all the sequences from three databases are pooled together.

Additionally, let us look at the case of H.264 streams subject to channel-induced

distortions with respect to $\mathbf{D}_k^{\text{SE}}$ operations in Table 19. In spatial domain operations, the LIVE database produced the `max` and `kurtosis`, the CSIQ database produced `mean to standard deviation ratio`, and finally the IVP database produced `mean to standard deviation ratio` and `max`. The cross-database results show that `max` and `kurtosis` are the most effective operation when all the sequences from three databases are pooled together. Furthermore, in the temporal domain, the variability of operations is higher. The LIVE database produced the `standard deviation` and `max`, the CSIQ database produced `standard deviation`, `skewness` and `kurtosis`, and finally the IVP database produced `mean` and l_2`-norm`. The cross-database results show that `mean` and `median` are the most effective operation when all the sequences from three databases are pooled together. In this case, the three databases are statistically uncorrelated when the sequences are processed separately.

This shows that the development process and subjective scores processing and regularization methods are different among the databases. This irregularity leads to ambiguous and misleading results when an approach is validated using the data of one database or another. This shows a need for a validation and a verification framework for perceptual quality assessment databases. The existence of such framework would guarantee a consistency in PVQA databases and lead the community to better understanding of human perception.

5.2.2.2 **Pooling Operations Performance**

Table 17 summarizes the number of occurrences of operations in Tables 19,23, 27. Each entry in this table, represents the number of occurrences of the pooling operation in the respective domain in Tables 19,23, 27. Furthermore, Table 18 summarizes the number of occurrences of these pooling operations in terms of PLOCC. We note here the following key observations from Tables 17-18.

- In the spatial domain, there is an overwhelming dominance of the `maximum` for

Table 17: Number of occurrences of video pooling operations among the best three scenarios of SROCC for each distortion map. The red cells indicate that the operation's frequency is 35% or higher, and the blue ones indicate 20%-35% frequency. These numbers are based on the statistics in Tables 19,23, 27.

Operation	Spatial				Temporal			
	SE	SSIM	OF	All Maps	SE	SSIM	OF	All Maps
Total Count	54	54	54	162	54	54	54	162
mean	1	4	2	7	**14**	7	**12**	**33**
std	4	**20**	4	28	6	5	6	17
mean/std	9	**15**	2	26	3	3	2	8
kurtosis	**12**	3	**15**	30	2	2	5	9
skewness	2	4	0	6	3	2	3	8
max	**24**	4	**28**	**56**	8	9	6	23
l_1-norm	1	3	1	5	5	4	1	10
l_2-norm	1	1	2	4	5	8	8	21
median					8	**14**	**11**	**33**

Table 18: Number of occurrences of video pooling operations among the best three scenarios of PLOCC for each distortion map. The red cells indicate that the operation's frequency is 35% or higher, and the blue ones indicate 20%-35% frequency. These numbers are based on the statistics in Tables 20,24, 28.

Operation	Spatial				Temporal			
	SE	SSIM	OF	All Maps	SE	SSIM	OF	All Maps
Total Count	54	54	54	162	54	54	54	162
mean	3	0	2	5	**16**	8	**8**	**32**
std	4	**28**	4	**36**	7	3	6	16
mean/std	3	**15**	3	21	6	4	1	11
kurtosis	**14**	3	**15**	**32**	2	1	7	10
skewness	1	1	1	3	3	2	3	8
max	**25**	6	**25**	**56**	7	7	6	20
l_1-norm	2	1	3	6	6	8	7	21
l_2-norm	**2**	0	**1**	3	3	**11**	**8**	22
median					4	**10**	**8**	22

the $\mathbf{D}_k^{\mathrm{SE}}$ and $\mathbf{D}_k^{\mathrm{OF}}$. The `maximum` was frequent in almost 46% of the operations in both maps. ***This shows that the human perception is mostly distracted by the maximum distortion in the spatial domain in terms of fidelity and motion distortion.***

- Furthermore, the `standard deviation` seems to dominate the spatial operations based on $\mathbf{D}_k^{\mathrm{SSIM}}$. This indicates that ***human perception is sensitive to the variability of structural distortion in the frames.*** However, the temporal pooling is highly variant for $\mathbf{D}_k^{\mathrm{SSIM}}$. The most effective temporal pooling operations in this case are `median`, `mean` and l_2-`norm`.

- In the temporal domain, the `mean` and `median` were dominant, which indicates that ***human perception is mostly affected by the average temporal variations in distortion.***

5.A Best Performing Pooling Strategies for the Tested Distortion Maps

Table 19: The best pooling strategies for full reference squared error distortion maps (spatial pooling, temporal pooling, SROCC, average frames per sequence).

Database	Distortion Type	Total Videos	Best	2nd	3rd
LIVE	Compression (H.264)	40	(max,max,0.926,407)	(kurtosis,max,0.926,407)	(max,mean,0.920,407)
	Channel induced (H.264)	40	**(max,std,0.896,407)**	**(kurtosis,std,0.896,407)**	(max,max,0.882,407)
	Frame-freezes (stored content)	30	**(max,max,0.909,449)**	**(kurtosis,max,0.909,449)**	(max,std,0.889,449)
	Frame-freezes (live feed)*	10	(mean/std,mean/std,0.612,271)	(skewness,median,0.600,271)	(skewness,mean,0.576,271)
	Rate Adaptation	27	(max,std,0.888,407)	(kurtosis,std,0.888,407)	(max,max,0.868,407)
	Temporal Dynamics*	50	(std,norm2,0.828,211)	(std,mean,0.821,211)	(std,norm1,0.821,211)
CSIQ	Compression (H.264)	36	(max,skewness,0.910,380)	(kurtosis,skewness,0.910,380)	(max,kurtosis,0.884,380)
	Compression (HEVC)	36	(max,mean/std,0.842,380)	(kurtosis,mean/std,0.842,380)	(std,median,0.833,380)
	Compression (MJPEG)	36	**(max,max,0.829,381)**	(kurtosis,max,0.829,381)	(max,median,0.825,381)
	Compression, wavelet based (SNOW)	36	(max,median,0.916,380)	(kurtosis,median,0.916,380)	(max,mean,0.915,380)
	Channel induced (H.264)	36	**(mean/std,skewness,0.928,380)**	(mean/std,kurtosis,0.925,380)	(mean/std,std,0.904,380)
	Additive white noise (AWGN)	36	(mean,mean,0.903,380)	(norm1,mean,0.903,380)	(norm2,median,0.901,380)
IVP	Compression (H.264)	40	**(max,mean,0.866,228)**	**(kurtosis,mean,0.866,228)**	**(max,norm2,0.866,228)**
	Compression (MPEG2)	30	(mean/std,norm1,0.710,229)	(mean/std,median,0.708,229)	(mean/std,norm2,0.700,229)
	Compression (Dirac)	30	(max,norm1,0.764,229)	(kurtosis,norm1,0.764,229)	(max,mean,0.762,229)
	Channel induced (H.264)	28	(mean/std,mean,0.672,219)	(mean/std,norm1,0.665,219)	(max,mean,0.659,219)
All	Channel induced (H.264)	104	(max,mean,0.721,347)	(kurtosis,mean,0.721,347)	(max,median,0.704,347)
	Compression (H.264)	116	(max,norm2,0.833,337)	(kurtosis,norm2,0.833,337)	(max,mean,0.832,337)

Table 20: The best pooling strategies for full reference squared error distortion maps (spatial pooling, temporal pooling, PLOCC, average frames per sequence).

Database	Distortion Type	Total Videos	Best	2nd	3rd
LIVE	Compression (H.264)	40	(max,max,0.884,407)	(kurtosis,max,0.884,407)	(max,mean,0.838,407)
	Channel induced (H.264)	40	**(max,std,0.869,407)**	**(kurtosis,std,0.869,407)**	(max,max,0.845,407)
	Frame-freezes (stored content)	30	(max,max,0.871,449)	(kurtosis,max,0.871,449)	(mean,kurtosis,0.862,449)
	Frame-freezes (live feed)*	10	**(skewness,kurtosis,0.620,271)**	(norm1,std,0.590,271)	(mean,std,0.590,271)
	Rate Adaptation	27	**(std,mean/std,0.884,449)**	(norm2,mean/std,0.881,449)	(max,mean/std,0.877,407)
	Temporal Dynamics*	50	(max,mean,0.821,211)	(kurtosis,mean,0.821,211)	(max,norm1,0.821,211)
CSIQ	Compression (H.264)	36	**(max,skewness,0.863,380)**	**(kurtosis,skewness,0.863,380)**	(std,mean,0.804,380)
	Compression (HEVC)	36	(max,mean/std,0.795,380)	(kurtosis,mean/std,0.795,380)	(std,median,0.767,380)
	Compression (MJPEG)	36	**(max,median,0.805,381)**	**(kurtosis,median,0.805,381)**	(max,max,0.804,381)
	Compression, wavelet based (SNOW)	36	(max,std,0.871,380)	(kurtosis,std,0.871,380)	(max,max,0.870,380)
	Channel induced (H.264)	36	**(mean/std,std,0.894,380)**	**(mean/std,skewness,0.894,380)**	(max,mean/std,0.820,380)
	Additive white noise (AWGN)	36	**(norm1,mean,0.843,380)**	**(mean,mean,0.843,380)**	(norm2,mean,0.837,380)
IVP	Compression (H.264)	40	(max,norm1,0.796,228)	(kurtosis,norm1,0.796,228)	(max,norm2,0.793,228)
	Compression (MPEG2)	30	(max,mean,0.728,229)	(kurtosis,mean,0.728,229)	(max,norm1,0.726,229)
	Compression (Dirac)	30	(max,norm1,0.743,229)	(kurtosis,norm1,0.743,229)	(max,mean,0.738,229)
	Channel induced (H.264)	28	(max,mean,0.670,219)	(kurtosis,mean,0.670,219)	(mean/std,mean,0.663,219)
All	Channel induced (H.264)	104	**(max,mean,0.699,347,104)**	**(kurtosis,mean,0.699,347,104)**	(max,median,0.663,347,104)
	Compression (H.264)	116	(std,mean,0.749,337)	(max,norm2,0.742,337)	(kurtosis,norm2,0.742,337)

Table 21: Spearman correlation coefficients statistical significance of the tested spatial pooling strategies using SE distortion map.

	(mean,mean)			(std,mean)			(mean/std,mean)			(norm1,mean)			(norm2,mean)			(max,mean)			(kurtosis,mean)			(skewness,mean)			Operation Total		
	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP
(mean,mean)	0	0	0	0	1	0	1	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0	1	0	1	4	2
(std,mean)	0	1	0	0	0	0	1	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0	1	0	1	6	0
(mean/std,mean)	1	1	0	1	1	0	0	0	0	1	1	0	1	1	0	1	1	0	1	1	0	0	0	0	6	6	0
(norm1,mean)	0	0	0	0	1	0	1	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0	1	0	1	4	2
(norm2,mean)	0	1	0	0	0	0	1	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0	1	0	1	6	0
(max,mean)	0	0	1	0	1	0	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	1	4	3
(kurtosis,mean)	0	0	1	0	1	0	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	1	4	3
(skewness,mean)	0	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0	1	1	0	1	1	0	0	0	0	6	2
Database Total	1	4	2	1	6	0	6	6	0	1	4	2	1	6	0	1	4	3	1	4	3	0	6	2	12	40	12
Databases Total	7			7			12			7			7			8			8			8			64		

Table 22: Spearman correlation coefficients statistical significance of the tested temporal pooling strategies using SE distortion map.

	(mean,mean)			(std,mean)			(mean/std,mean)			(norm1,mean)			(norm2,mean)			(max,mean)			(median,mean)			(kurtosis,mean)			(skewness,mean)			Operation Total		
	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP
(max,mean)	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	0	1	5	2
(max,std)	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	3	1
(max,mean/std)	0	1	1	0	1	1	0	0	0	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	0	1	0	7	8
(max,norm1)	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	2	1
(max,norm2)	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	2	2
(max,max)	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	3	1
(max,median)	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	5	1
(max,kurtosis)	0	1	1	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	3	3
(max,skewness)	1	1	0	1	0	0	0	0	1	1	1	0	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	5	4	1
Database Total	1	5	2	1	3	1	0	7	8	1	2	1	1	2	2	1	3	1	0	5	1	0	3	3	5	4	1	10	34	20
Databases Total	8			5			15			4			5			5			6			6			10			64		

Table 23: The best pooling strategies for full reference SSIM distortion maps (spatial pooling, temporal pooling, SROCC, average frames per sequence).

Database	Distortion Type	Total Videos	Best	2nd	3rd
LIVE	Compression (H.264)	40	(std,max,0.836,407)	(std,std,0.835,407)	(std,norm2,0.828,407)
	Channel induced (H.264)	40	(std,std,0.854,407)	(std,mean,0.849,407)	(std,norm1,0.849,407)
	Frame-freezes (stored content)	30	(mean/std,mean/std,0.907,449)	(max,max,0.906,449)	(kurtosis,max,0.906,449)
	Frame-freezes (live feed)*	10	(mean/std,kurtosis,0.552,271)	(skewness,norm2,0.442,271)	(skewness,mean,0.430,271)
	Rate Adaptation	27	(std,std,0.812,407)	(std,max,0.795,407)	(std,norm2,0.794,407)
	Temporal Dynamics*	50	**(std,mean,0.840,211)**	**(std,norm1,0.840,211)**	(std,norm2,0.837,211)
CSIQ	Compression (H.264)	36	**(mean,median,0.939,380)**	**(norm1,median,0.939,380)**	**(norm2,median,0.939,380)**
	Compression (HEVC)	36	**(mean/std,norm2,0.937,380)**	(mean/std,max,0.922,380)	(mean,median,0.916,380)
	Compression (MJPEG)	36	(skewness,skewness,0.200,381)	(max,mean/std,0.190,381)	(kurtosis,mean/std,0.190,381)
	Compression, wavelet based (SNOW)	36	**(mean/std,norm2,0.924,380)**	(mean/std,max,0.922,380)	(mean/std,mean,0.917,380)
	Channel induced (H.264)	36	(mean/std,kurtosis,0.913,380)	(skewness,std,0.912,380)	(mean/std,skewness,0.912,380)
	Additive white noise (AWGN)	36	**(max,max,0.906,380)**	**(kurtosis,max,0.906,380)**	(max,median,0.903,380)
IVP	Compression (H.264)	40	(std,median,0.852,228)	(std,norm1,0.845,228)	(std,mean,0.839,228)
	Compression (MPEG2)	30	(mean/std,max,0.843,229)	(mean/std,std,0.812,229)	(mean/std,norm2,0.800,229)
	Compression (Dirac)	30	(mean/std,mean,0.761,229)	(mean,median,0.759,229)	(norm1,median,0.759,229)
	Channel induced (H.264)	28	(mean/std,median,0.660,219)	(mean,median,0.622,219)	(norm1,median,0.622,219)
All	Channel induced (H.264)	104	(std,median,0.665,347)	(mean/std,median,0.624,347)	(std,norm1,0.614,347)
	Compression (H.264)	116	**(std,mean,0.862,337)**	(std,norm2,0.857,337)	(std,median,0.850,337)

Table 24: The best pooling strategies for full reference SSIM distortion maps (spatial pooling, temporal pooling, PLOCC, average frames per sequence).

Database	Distortion Type	Total Videos	Best	2nd	3rd
LIVE	Compression (H.264)	40	(std,max,0.831,407)	(std,norm2,0.807,407)	(std,mean,0.807,407)
	Channel induced (H.264)	40	(std,norm2,0.813,407)	(std,mean,0.813,407)	(std,norm1,0.813,407)
	Frame-freezes (stored content)	30	**(std,kurtosis,0.898,449)**	(mean/std,mean/std,0.865,449)	(norm1,median,0.855,449)
	Frame-freezes (live feed)*	10	(max,skewness,0.492,271)	(kurtosis,skewness,0.492,271)	(max,mean/std,0.491,271)
	Rate Adaptation	27	(mean/std,mean/std,0.816,449)	(std,max,0.776,407)	(std,std,0.774,407)
	Temporal Dynamics*	50	**(std,norm2,0.840,211)**	(std,norm1,0.838,211)	(std,mean,0.838,211)
CSIQ	Compression (H.264)	36	(mean/std,norm1,0.806,380)	(std,norm2,0.766,380)	(mean/std,median,0.760,380)
	Compression (HEVC)	36	**(mean/std,max,0.922,380)**	(mean/std,norm2,0.895,380)	(mean/std,std,0.885,380)
	Compression (MJPEG)	36	(max,median,0.231,381)	(kurtosis,median,0.231,381)	(max,max,0.190,381)
	Compression, wavelet based (SNOW)	36	**(mean/std,max,0.893,380)**	(std,median,0.889,380)	(std,max,0.888,380)
	Channel induced (H.264)	36	(skewness,std,0.874,380)	(mean/std,mean,0.865,380)	(mean/std,mean/std,0.861,380)
	Additive white noise (AWGN)	36	(max,median,0.813,380)	(kurtosis,median,0.813,380)	(max,max,0.774,380)
IVP	Compression (H.264)	40	(std,norm1,0.829,228)	(std,mean,0.819,228)	(std,norm2,0.814,228)
	Compression (MPEG2)	30	(std,norm2,0.787,229)	(std,norm1,0.783,229)	(std,mean,0.775,229)
	Compression (Dirac)	30	(mean/std,mean,0.798,229)	(mean/std,median,0.793,229)	(mean/std,norm2,0.790,229)
	Channel induced (H.264)	28	(mean/std,median,0.646,219)	(std,norm2,0.439,219)	(std,norm1,0.430,219)
All	Channel induced (H.264)	104	(mean/std,median,0.425,347)	(std,norm1,0.397,347)	(std,norm2,0.392,347)
	Compression (H.264)	116	(std,mean,0.729,337)	(std,norm2,0.723,337)	(std,norm1,0.711,337)

Table 25: Spearman correlation coefficients statistical significance of the tested spatial pooling strategies using SSIM distortion map.

	(mean,mean)			(std,mean)			(mean/std,mean)			(norm1,mean)			(norm2,mean)			(max,mean)			(kurtosis,mean)			(skewness,mean)			Operation Total		
	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP
(mean,mean)	0	0	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	2	2
(std,mean)	1	1	1	0	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	6	6	6
(mean/std,mean)	0	1	1	1	0	0	0	0	0	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	2	6	6
(norm1,mean)	0	0	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	2	2
(norm2,mean)	0	0	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	2	2
(max,mean)	0	0	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	3
(kurtosis,mean)	0	0	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	3
(skewness,mean)	1	0	0	0	1	1	1	1	1	1	0	0	1	0	0	0	0	1	0	0	1	0	0	0	4	2	4
Database Total	2	2	2	6	6	6	2	6	6	2	2	2	2	2	2	1	2	3	1	2	3	4	2	4	20	24	28
Databases Total	6			18			14			6			6			6			6			10			72		

Table 26: Spearman correlation coefficients statistical significance of the tested temporal pooling strategies using SSIM distortion map.

	(mean,mean)			(std,mean)			(mean/std,mean)			(norm1,mean)			(norm2,mean)			(max,mean)			(median,mean)			(kurtosis,mean)			(skewness,mean)			Operation Total		
	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP
(max,mean)	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
(max,std)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	2	0
(max,mean/std)	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	5	0	0
(max,norm1)	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
(max,norm2)	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
(max,max)	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
(max,median)	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
(max,kurtosis)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(max,skewness)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Database Total	1	0	0	0	2	0	5	0	0	1	0	0	1	0	0	1	1	0	1	1	0	0	0	0	0	0	0	10	4	0
Databases Total	1			2			5			1			1			2			2			0			0			14		

Table 27: The best pooling strategies for full reference optical flow diffrence distortion maps (spatial pooling, temporal pooling, SROCC, average frames per sequence).

Database	Distortion Type	Total Videos	Best	2nd	3rd
LIVE	Compression (H.264)	40	**(max,max,0.933,449)**	**(kurtosis,max,0.933,449)**	(max,norm2,0.922,449)
	Channel induced (H.264)	40	(max,norm2,0.863,449)	(kurtosis,norm2,0.863,449)	(max,mean,0.862,449)
	Frame-freezes (stored content)	30	(max,max,0.882,449)	(kurtosis,max,0.882,449)	(max,kurtosis,0.882,449)
	Frame-freezes (live feed)*	10	**(norm2,kurtosis,0.842,271)**	(std,kurtosis,0.782,271)	(norm2,skewness,0.745,271)
	Rate Adaptation	27	**(max,kurtosis,0.890,449)**	**(kurtosis,kurtosis,0.890,449)**	(max,std,0.869,449)
	Temporal Dynamics*	50	(max,norm2,0.771,211)	(kurtosis,norm2,0.771,211)	(max,mean,0.768,211)
CSIQ	Compression (H.264)	36	(max,median,0.902,381)	(kurtosis,median,0.902,381)	(max,mean,0.886,381)
	Compression (HEVC)	36	(max,std,0.865,381)	(kurtosis,std,0.865,381)	(max,max,0.855,381)
	Compression (MJPEG)	36	(max,skewness,0.579,381)	(kurtosis,skewness,0.579,381)	(std,mean,0.477,381)
	Compression, wavelet based (SNOW)	36	(max,std,0.832,381)	(kurtosis,std,0.832,381)	(max,max,0.757,381)
	Channel induced (H.264)	36	(mean/std,std,0.894,381)	(mean/std,mean/std,0.883,381)	(max,mean/std,0.869,381)
	Additive white noise (AWGN)	36	(mean,median,0.694,381)	(norm1,median,0.694,381)	(mean,mean,0.676,381)
IVP	Compression (H.264)	40	(max,mean,0.847,253)	(kurtosis,mean,0.847,253)	(max,norm1,0.846,253)
	Compression (MPEG2)	30	**(max,median,0.879,253)**	**(kurtosis,median,0.879,253)**	(max,mean,0.869,253)
	Compression (Dirac)	30	**(max,median,0.894,253)**	**(kurtosis,median,0.894,253)**	(max,mean,0.887,253)
	Channel induced (H.264)	28	**(max,median,0.818,254)**	**(kurtosis,median,0.818,254)**	(std,norm2,0.799,254)
All	Channel induced (H.264)	104	**(max,mean,0.773,373)**	**(kurtosis,mean,0.773,373)**	(max,median,0.768,373)
	Compression (H.264)	116	(max,norm2,0.797,360)	(kurtosis,norm2,0.797,360)	(std,mean,0.794,360)

Table 28: The best pooling strategies for full reference optical flow difference distortion maps (spatial pooling, temporal pooling, PLOCC, average frames per sequence).

Database	Distortion Type	Total Videos	Best	2nd	3rd
LIVE	Compression (H.264)	40	**(max,max,0.884,449)**	**(kurtosis,max,0.884,449)**	(max,norm2,0.877,449)
	Channel induced (H.264)	40	(max,norm2,0.840,449)	(kurtosis,norm2,0.840,449)	(max,mean,0.839,449)
	Frame-freezes (stored content)	30	(max,kurtosis,0.882,449)	(kurtosis,kurtosis,0.882,449)	(std,kurtosis,0.869,449)
	Frame-freezes (live feed)*	10	(mean,max,0.470,271)	(norm1,max,0.470,271)	(std,kurtosis,0.463,271)
	Rate Adaptation	27	(max,kurtosis,0.862,449)	(kurtosis,kurtosis,0.862,449)	(norm2,kurtosis,0.828,449)
	Temporal Dynamics*	50	(max,mean,0.790,211)	(kurtosis,mean,0.790,211)	(max,norm1,0.790,211)
CSIQ	Compression (H.264)	36	(max,median,0.859,381)	(kurtosis,median,0.859,381)	(max,mean,0.854,381)
	Compression (HEVC)	36	(max,std,0.836,381)	(kurtosis,std,0.836,381)	(max,max,0.807,381)
	Compression (MJPEG)	36	(max,skewness,0.562,381)	(kurtosis,skewness,0.562,381)	(std,median,0.451,381)
	Compression, wavelet based (SNOW)	36	(max,std,0.833,381)	(kurtosis,std,0.833,381)	(max,max,0.783,381)
	Channel induced (H.264)	36	(mean/std,std,0.873,381)	(skewness,std,0.865,381)	(mean/std,mean/std,0.849,381)
	Additive white noise (AWGN)	36	(norm1,median,0.720,381)	(mean,median,0.720,381)	(norm1,mean,0.694,381)
IVP	Compression (H.264)	40	**(max,norm1,0.855,253)**	**(kurtosis,norm1,0.855,253)**	(max,norm2,0.843,253)
	Compression (MPEG2)	30	**(max,median,0.865,253)**	**(kurtosis,median,0.865,253)**	**(max,mean,0.865,253)**
	Compression (Dirac)	30	**(max,norm1,0.866,253)**	**(kurtosis,norm1,0.866,253)**	(max,median,0.864,253)
	Channel induced (H.264)	28	**(mean/std,skewness,0.783,254)**	(max,norm1,0.737,254)	(kurtosis,norm1,0.737,254)
All	Channel induced (H.264)	104	(max,norm2,0.670,373)	(kurtosis,norm2,0.670,373)	(max,mean,0.665,373)
	Compression (H.264)	116	**(max,norm2,0.787,360)**	**(kurtosis,norm2,0.787,360)**	(std,mean,0.777,360)

Table 29: Spearman correlation coefficients statistical significance of the tested spatial pooling strategies using optical flow difference distortion map.

	(mean,mean)			(std,mean)			(mean/std,mean)			(norm1,mean)			(norm2,mean)			(max,mean)			(kurtosis,mean)			(skewness,mean)			Operation Total		
	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP
(mean,mean)	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0	0	1	0	0	1	1	1	0	2	2	4
(std,mean)	0	0	1	0	0	0	1	1	1	0	0	1	0	0	0	0	0	1	0	0	1	1	1	1	2	2	6
(mean/std,mean)	1	1	0	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	6	6	4
(norm1,mean)	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0	0	1	0	0	1	1	1	0	2	2	4
(norm2,mean)	0	0	1	0	0	0	1	1	1	0	0	1	0	0	0	0	0	1	0	0	1	1	1	1	2	2	6
(max,mean)	0	0	1	0	0	1	1	1	1	0	0	1	0	0	1	0	0	0	0	0	0	1	1	1	2	2	6
(kurtosis,mean)	0	0	1	0	0	1	1	1	1	0	0	1	0	0	1	0	0	0	0	0	0	1	1	1	2	2	6
(skewness,mean)	1	1	0	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	6	6	4
Database Total	2	2	4	2	2	6	6	6	4	2	2	4	2	2	6	2	2	6	2	2	6	6	6	4	24	24	40
Databases Total	8			10			16			8			10			10			10			16			88		

Table 30: Spearman correlation coefficients statistical significance of the tested temporal pooling strategies using optical flow difference distortion map.

	(mean,mean)			(std,mean)			(mean/std,mean)			(norm1,mean)			(norm2,mean)			(max,mean)			(median,mean)			(kurtosis,mean)			(skewness,mean)			Operation Total		
	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP	LIVE	CSIQ	IVP
(max,mean)	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	2	3
(max,std)	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	2	3
(max,mean/std)	0	1	1	0	1	1	0	0	0	0	1	1	0	1	1	0	1	1	0	1	1	0	1	0	0	0	1	0	7	7
(max,norm1)	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	2	3
(max,norm2)	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	2	3
(max,max)	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	2	3
(max,median)	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	2	3
(max,kurtosis)	0	0	1	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	0	1	1	0	2	7
(max,skewness)	0	1	1	0	1	1	0	0	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	0	0	0	7	8
Database Total	0	2	3	0	2	3	0	7	7	0	2	3	0	2	3	0	2	3	0	2	3	0	2	7	0	7	8	0	28	40
Databases Total	5			5			14			5			5			5			5			9			15			68		

CHAPTER VI

CONCLUSION AND FUTURE DIRECTIONS

This dissertation addresses perceptual video quality assessment in different applications, contents and conditions. We started by introducing a comprehensive background about video technology, coding, and perceptual quality assessment. We tie this work and closely examine the correlation with visual perception and the characteristic of the human visual system. The contributions of this thesis are three-fold. Firstly, we propose PeQASO, a perceptual video quality assessment approach using optical flow-based distortions maps. We propose a reduced-reference perceptual video quality metric to estimate distortion due to compression and network losses. The proposed technique does not make any assumption about the coding conditions or video sequence. It rather explores the temporal changes between the frames by analyzing the variations in the statistical properties of the optical flow. We validate our approach by testing it on various sequences and compare our estimated quality metric with the DMOS values at the sequences level reported in three independent databases for sequences subject to network errors or losses. Our experiments show that the proposed technique captures the perceptual quality very well. PeQASO was the first work to utilize pixel-level optical flow maps to examine the fidelity of motion fields.

Secondly, we propose utilizing power spectral analysis to estimate the perceptual quality of videos. We start by designing a low-complexity no-reference video quality measure to estimate the channel-induced distortion at the frame-level due to network losses. More importantly, we propose POTUS, a perceptual objective quality

assessment framework based on tempospatially unified power-spectral density characteristics. POTUS is a perceptual video quality assessment metric for distorted videos by analyzing the power spectral density of a group of pictures. This is an estimation approach that relies on the changes in video dynamic calculated in the frequency domain and are primarily caused by distortion. We obtain a feature map by processing a 3D PSD tensor obtained from a set of distorted frames. This is a full-reference tempospatial approach that considers both temporal and spatial PSD characteristics. This makes it ubiquitously suitable for videos with different motion patterns and spatial contents. Our technique does not make any assumptions on the coding conditions, streaming conditions or distortion. This approach is also computationally inexpensive which makes it feasible for real-time and practical implementations. We validate our proposed metric by testing it on a variety of distorted sequences from PVQA databases. The results show that our metric estimates the perceptual quality at the sequence level accurately. We report the correlation coefficients with the differential mean opinion scores reported in the databases. The results show high and competitive correlations compared with the state of the art techniques. This is also the first time the power spectral density have been utilized as a feature to understand and evaluate video quality.

Finally, we analyze three distortion maps spatially and temporally, and identify the most effective statistical moments and pooling strategies with respect to PVQA. The three distortion maps examine three visual feature: pixel fidelity, local structural similarity and motion fields. We show the most significant spatial and temporal features correlated with perception for every distortion map. For every distortion type in the databases, we identify the best performing operation for each map and the overall best. We use this data to draw insights about the human perception and its sensitivity to distortion. We also demonstrate that the same distortions across databases yield different results in terms of PVQA evaluation and verification. This

warrants the need for a verification and validation framework for PVQA databases.

6.1 Future Research Directions

Perceptual video quality remains a highly emphasized field of research. We believe that deeper understanding of the human visual processing is an essential ingredient to design accurate and practical visual computational models. This work will continue investigating various human visual processing characteristics that can be incorporated in future video systems. The processing mechanism of the HVS remains an open field of investigation in computational neuroscience. As future studies and research continue to reveal more about the HVS, especially temporal and motion processing, we will continue to examine these findings and incorporate them in visual media processing systems.

Additionally, the work proposed herein can be exploited to serve purposes beyond perceptual video quality. The proposed features and algorithms can be utilized in several video applications including video retrieval, classification and enhancement. Video quality assessment algorithms work as a measure of the variation between video contents in different features domains. Hence, such algorithms can be extended to discriminate between features of different video classes and contents. Furthermore, the inherited perceptual nature of video quality algorithms makes them suitable to be used as optimization engines for video enhancement algorithms. We plan to explore the utility of the proposed features and metrics in such applications.

Furthermore, deep learning can be utilized in this work to facilitate more applications and efficient processing of the proposed techniques. We plan to explore architecture design and the utility of deep learning networks in video quality assessment applications. The proposed full- and reduced-reference metrics can be extended using deep learning to facilitate real-time no-reference video quality monitoring. The handcrafted features and distortion maps can be incorporated into the deep network

design to efficiently and accurately serve diverse applications.

Bibliography

[1] P. Research, *Parabola Explorer Software, Version 2.5*, University of Southampton Science Park, UK, 2013.

[2] G. J. Sullivan, "Overview of international video coding standards (preceding h.264/avc)," https://www.itu.int/ITU-T/worksem/vica/docs/presentations/S0_P2_Sullivan.pdf, July 2005, (Accessed on 10/19/2016).

[3] "Alliance for open media - aom in the news," http://aomedia.org/aom-in-the-news/, (Accessed on 10/20/2016).

[4] J. Loomis and M. Wasson, "Vc-1 technical overview," https://www.microsoft.com/windows/windowsmedia/howto/articles/vc1techoverview.aspx, October 2007, microsoft Corporation (Accessed on 10/17/2016).

[5] "Audio video coding standard workgroup of china," http://www.avs.org.cn/english/Achievement.asp#2.1, (Accessed on 10/20/2016).

[6] X.-F. Wang and D.-B. Zhao, "Performance comparison of avs and h.264/avc video coding standards," *Journal of Computer Science and Technology*, vol. 21, no. 3, pp. 310–314, 2006.

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

[8] I. K. Kim, S. Lee, Y. Piao, and J. Chen, "Coding efficiency comparison of new video coding standards: Hevc vs vp9 vs avs2 video," in *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on*, July 2014, pp. 1–6.

[9] M. Rerabek and T. Ebrahimi, "Comparison of compression efficiency between hevc/h.265 and vp9 based on subjective assessments," in *Proc. SPIE*, vol. 9217, 2014, pp. 92 170U–13.

[10] "The webm project — vp9 video codec summary," https://www.webmproject.org/vp9/, (Accessed on 10/20/2016).

[11] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h.264/avc video coding standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 560–576, July 2003.

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

[13] C. J. Perry and M. Fallah, "Feature integration and object representations along the dorsal stream visual hierarchy," *Frontiers in Computational Neuroscience*, vol. 8, p. 84, 2014.

[14] ITU-T, "P.910: Subjective video quality assessment methods for multimedia applications," ITU Telecommunication Standardization Sector, Tech. Rep., 2008.

[15] P. V. Vu and D. M. Chandler, "Vis3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *Journal of Electronic Imaging*, vol. 23, no. 1, p. 013016, 2014.

[16] "Cisco visual networking index (vni) global mobile data traffic forecast," https://www.ciscoknowledgenetwork.com/files/573_02-23-16-Documents2016_VNI_Mobile_CKN_Final.pdf?PRIORITY_CODE=194542_20, Feb 2016, accessed: 2016-12-01.

[17] "The zettabyte era: Trends and analysis," http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.pdf, Feb 2016, accessed: 2014-12-01.

[18] "How consumers judge their viewing experience," http://lp.conviva.com/rs/901-ZND-194/images/CSR2015_HowConsumersJudgeTheirViewingExperience_Final.pdf, 2015, (Accessed on 11/28/2016).

[19] F. Bossen, "Common test conditions and software reference configurations," in *Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCTVC-K1100*, Shanghai, China, 11th meeting, oct 2012.

[20] F. Bossen, D. Flynn, and K. Sühring, *HM 12.1 Software Manual*, Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCTVC-Software Manual, may 2013.

[21] M. Naccari, M. Tagliasacchi, and S. Tubaro, "No-reference video quality monitoring for h.264/avc coded video," *Multimedia, IEEE Transactions on*, vol. 11, no. 5, pp. 932–946, Aug 2009.

[22] G. Valenzise, S. Magni, M. Tagliasacchi, and S. Tubaro, "No-reference pixel video quality monitoring of channel-induced distortion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 4, pp. 605–618, April 2012.

[23] A. C. Bovik, *The Essential Guide to Video Processing*, 2nd ed. Academic Press, 2009.

[24] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 688–703, July 2003.

[25] B. Bross, W.-J. Han, J.-R. Ohm, G. J. Sullivan, Y.-K. Wang, and T. Wiegand, "High efficiency video coding (hevc) text specification draft 10 (for fdis & final call)," Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCTVC-L1003.v34, Tech. Rep., jan 2013.

[26] I. Richardson and A. Bhat. Historical timeline of video coding standards and formats - vcodex. https://www.vcodex.com/historical-timeline-of-video-coding-standards-and-formats/.

[27] A. M. Tekalp, *Digital Video Processing*, 2nd ed. Prentice Hall, 2015.

[28] ITU, "Codecs for videoconferencing using primary digital group transmission," The International Telegraph and Telephone Consultative Committee (CCITT), Tech. Rep., 1988.

[29] ITU-T, "Advanced video coding for generic audiovisual services - recommendation 03/2005," ITU Telecommunication Standardization Sector, Tech. Rep., 2005.

[30] ——, "Advanced video coding for generic audiovisual services - recommendation 11/2007," ITU Telecommunication Standardization Sector, Tech. Rep., 2007.

[31] ——, "Advanced video coding for generic audiovisual services - recommendation 03/2009," ITU Telecommunication Standardization Sector, Tech. Rep., 2009.

[32] ——, "Advanced video coding for generic audiovisual services - recommendation 02/2014," ITU Telecommunication Standardization Sector, Tech. Rep., 2014.

[33] , "2016 global media format report," Encoding.com, techreport, 2016, http://1yy04i3k9fyt3vqjsf2mv610yvm-wpengine.netdna-ssl.com/files/2016-Global-Media-Formats-Report.pdf (Accessed: November 2016).

[34] G. J. Sullivan, "Meeting report for 26th vceg meeting," ITU Telecommunication Standardization Sector, Tech. Rep., 2005.

[35] D. Grois, D. Marpe, T. Nguyen, and O. Hadar, "Comparative assessment of h.265/mpeg-hevc, vp9, and h.264/mpeg-avc encoders for low-delay video applications," in *Proc. SPIE*, vol. 9217, 2014, pp. 92 170Q–92 170Q–10.

[36] W. Gao and S. Ma, *Advanced Video Coding Systems*. Springer International Publishing, 2014.

[37] "The state of video codecs 2016 - streaming media magazine," http://www.streamingmedia.com/Articles/Editorial/Featured-Articles/The-State-of-Video-Codecs-2016-110117.aspx, 2016, (Accessed on 10/29/2016).

[38] "Matroska media container — matroska," https://www.matroska.org/, (Accessed on 11/03/2016).

[39] "Vp3 tutorial at it1me.com — get the facts, watch videos and explore vp3," http://www.it1me.com/learn?s=VP3, (Accessed on 11/03/2016).

[40] "Ietf begins standardization process for next-generation 'netvc' video codec (daala)," http://www.tomshardware.com/news/ietf-standardizes-netvc-daala-codec,28821.html, (Accessed on 11/03/2016).

[41] "Thor_2015-07-22-mo-v2.pptx," https://www.ietf.org/proceedings/93/slides/slides-93-netvc-4.pdf, (Accessed on 11/04/2016).

[42] "World, meet thor a project to hammer out a royalty free video codec," http://blogs.cisco.com/collaboration/world-meet-thor-a-project-to-hammer-out-a-royalty-free-video-codec, (Accessed on 11/04/2016).

[43] "What is av1? - streaming media magazine," http://www.streamingmedia.com/Articles/Editorial/What-Is-.../What-is-AV1-111497.aspx, (Accessed on 11/05/2016).

[44] M. Yuen and H. R. Wu, "A survey of hybrid mc/dpcm/dct video coding distortions," *Signal Process.*, vol. 70, no. 3, pp. 247–278, Nov. 1998.

[45] K. Zeng, T. Zhao, A. Rehman, and Z. Wang, "Characterizing perceptual artifacts in compressed video streams," in *Proc. SPIE*, vol. 9014, 2014, pp. 90 140Q–90 140Q–10.

[46] D. J. Kravitz, K. S. Saleem, C. I. Baker, L. G. Ungerleider, and M. Mishkin, "The ventral visual pathway: An expanded neural framework for the processing of object quality," *Trends in cognitive sciences*, vol. 17, no. 1, p. 2649, Jan 2013.

[47] D. Purves, G. J. Augustine, D. Fitzpatrick, L. C. Katz, A.-S. LaMantia, J. O. McNamara, and S. M. Williams, Eds., *Neuroscience - Types of Eye Movements and Their Functions*, 2nd ed. Sunderland (MA): Sinauer Associates, 2001, (Accessed on 11/10/2016).

[48] T. Lindeberg, "A computational theory of visual receptive fields," *Biological Cybernetics*, vol. 107, no. 6, pp. 589–635, 2013.

[49] D. W. Dong and J. J. Atick, "Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus," *Network: Computation in Neural Systems*, vol. 6, no. 2, pp. 159–178, 1995.

[50] J. H. Elder and A. J. Sachs, "Psychophysical receptive fields of edge detection mechanisms," *Vision Research*, vol. 44, no. 8, pp. 795 – 813, 2004.

[51] M. M. Murray, A. Thelen, G. Thut, V. Romei, R. Martuzzi, and P. J. Matusz, "The multisensory function of the human primary visual cortex," *Neuropsychologia*, vol. 83, pp. 161 – 169, 2016, special Issue: Functional Selectivity in Perceptual and Cognitive Systems - A Tribute to Shlomo Bentin (1946-2012).

[52] K. Seshadrinathan and A. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *Image Processing, IEEE Transactions on*, vol. 19, no. 2, pp. 335–350, Feb 2010.

[53] A. A. Ghazanfar and C. E. Schroeder, "Is neocortex essentially multisensory?" *Trends in Cognitive Sciences*, vol. 10, no. 6, pp. 278 – 285, 2006.

[54] M. S. Moore, "Psychophysical measurement and prediction of digital video quality," Ph.D. dissertation, University of California Santa Barbara, 2002.

[55] H. Strasburger, I. Rentschler, and M. Jttner, "Peripheral vision and pattern recognition: A review," *Journal of Vision*, vol. 11, no. 5, p. 13, 2011.

[56] M. K. Mandal, *Multimedia Signals and Systems*. Springer US, 2003.

[57] J. Mannos and D. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE Trans. Inf. Theor.*, vol. 20, no. 4, pp. 525–536, Sep. 2006.

[58] M. Naccari and M. Mrak, "Chapter 5 - perceptually optimized video compression," in *Academic Press Library in signal ProcessingImage and Video Compression and Multimedia*, ser. Academic Press Library in Signal Processing, S. Theodoridis and R. Chellappa, Eds. Elsevier, 2014, vol. 5, pp. 155 – 196.

[59] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 3, no. 2, pp. 253–265, April 2009.

[60] S. Gilaie-Dotan, A. P. Saygin, L. J. Lorenzi, R. Egan, G. Rees, and M. Behrmann, "The role of human ventral visual cortex in motion perception," *Brain*, vol. 136, no. 9, pp. 2784–2798, 2013.

[61] V. Q. E. G. (VQEG), "Final report from the video quality experts group on the validation of objective models of video quality assessment," Video Quality Experts Group, Tech. Rep., april 2000.

[62] T. Liu, Y. Wang, J. Boyce, H. Yang, and Z. Wu, "A novel video quality metric for low bit-rate video considering both coding and packet-loss artifacts," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 3, no. 2, pp. 280–293, April 2009.

[63] T. Liu, Y. Wang, J. Boyce, Z. Wu, and H. Yang, "Subjective quality evaluation of decoded video in the presence of packet losses," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1, April 2007, pp. I–1125–I–1128.

[64] X. Feng, T. Liu, D. Yang, and Y. Wang, "Saliency based objective quality assessment of decoded video affected by packet losses," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, Oct 2008, pp. 2560–2563.

[65] Y.-F. Ou, T. Liu, Z. Zhao, Z. Ma, and Y. Wang, "Modeling the impact of frame rate on perceptual quality of video," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, Oct 2008, pp. 689–692.

[66] 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, Sep. 2008, p. 6.

[67] Y.-F. Ou, Z. Ma, and Y. Wang, "A novel quality metric for compressed video considering both frame rate and quantization artifacts," in *International Workshop on Image Processing and Quality Metrics for Consumer (VPQM'08)*, vol. 80, Jan 2009, p. 100.

[68] F. Boulos, W. Chen, B. Parrein, and P. Le Callet, "Region-of-Interest Intra Prediction for H.264/AVC Error Resilience," in *IEEE International Conference on Image Processing*, Cairo, Egypt, Nov. 2009, pp. 3109 – 3112.

[69] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective assessment of h.264/avc video sequences transmitted over a noisy channel," in *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*, July 2009, pp. 204–209.

[70] C. Chen, L. K. Choi, G. de Veciana, C. Caramanis, R. Heath, and A. Bovik, "Modeling the time-varying subjective quality of http video streams with rate adaptations," *Image Processing, IEEE Transactions on*, vol. 23, no. 5, pp. 2206–2221, May 2014.

[71] A. Moorthy, L. K. Choi, A. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, Oct 2012.

[72] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, "Study of subjective and objective quality assessment of video," *Image Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1427–1441, June 2010.

[73] J.-S. Lee, F. De Simone, N. Ramzan, Z. Zhao, E. Kurutepe, T. Sikora, J. Ostermann, E. Izquierdo, and T. Ebrahimi, "Subjective evaluation of scalable video coding for content distribution," in *Proceedings of the International Conference on Multimedia*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 65–72.

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

[75] L. Goldmann, F. De Simone, and T. Ebrahimi, "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," in *Proc. SPIE, Three-Dimensional Image Processing (3DIP) and Applications*, vol. 7526, 2010, pp. 75 260S–75 260S–11.

[76] Y.-F. Ou, Y. Zhou, and Y. Wang, "Perceptual quality of video with frame rate variation: A subjective study," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 2446–2449.

[77] V. Q. E. G. (VQEG), "Report on the validation of video quality models for high definition video content," Video Quality Experts Group, Tech. Rep., april 2010.

[78] Y.-F. Ou, Z. Ma, T. Liu, and Y. Wang, "Perceptual quality assessment of video considering both frame rate and quantization artifacts," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 21, no. 3, pp. 286–298, March 2011.

[79] F. ZHANG, S. LI, L. MA, Y. C. WONG, and K. N. NGAN, "Ivp subjective quality video database," 2011, 00002.

[80] D. Chandler and S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *Image Processing, IEEE Transactions on*, vol. 16, no. 9, pp. 2284–2298, Sept 2007.

[81] E. Bosc, P. Le Callet, L. Morin, and M. Pressigout, "Visual quality assessment of synthesized views in the context of 3d-tv," in *3D-TV System with Depth-Image-Based Rendering Architectures, Techniques and Challenges*. Springer, 2012, pp. 439–474.

[82] Y. Pitrey, M. Barkowsky, P. Le Callet, and R. Pépion, "SUBJECTIVE QUALITY ASSESSMENT OF MPEG-4 SCALABLE VIDEO CODING IN A MOBILE SCENARIO," in *Second European Workshop on Visual Information Processing*, Paris, France, Jul. 2010, p. paper 72.

[83] ——, "Subjective Quality Evaluation of H.264 High-Definition Video Coding versus Spatial Up-Scaling and Interlacing," in *Euro ITV*, Tampere, Finland, Jun. 2010, p. irccyn contribution.

[84] Y. Pitrey, M. Barkowsky, R. Pépion, P. Le Callet, and H. Hlavacs, "Influence of the source content and encoding configuration on the perceived quality for scalable video coding," in *SPIE Human Vision and Electronic Imaging XVII*, vol. 8291, no. 54, San franscisco, United States, Jan. 2012, pp. 1–6.

[85] M. Barkowsky, N. Staelens, L. Janowski, Y. Koudota, M. Leszczuk, M. Urvoy, P. Hummelbrunner, I. Sedano, and K. Brunnstrm, "Subjective experiment dataset for joint development of hybrid video quality measurement algorithms," in *QoEMCS 2012 - Third Workshop on Quality of Experience for Multimedia Content Sharing*, Berlin, Germany, Jul 2012, pp. 1–4.

[86] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricordel, P. Le Callet, J. Gutierrez, and N. Garcia, "Nama3ds1-cospad1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3d stereoscopic sequences," in *QoMEX - Fourth International Workshop on Quality of Multimedia Experience*, Yarra Valley, Australia, Jul 2012, pp. 1–6.

[87] S. Péchard, M. Carnec, P. Le Callet, and D. Barba, "From SD to HD television: effects of H.264 distortions versus display size on quality of experience," in *International Conference on Image Processing*, Atlanta, United States, Oct. 2006, pp. 409–412.

[88] Y. Pitrey, U. Engelke, M. Barkowsky, R. Pépion, and P. Le Callet, "Aligning subjective tests using a low cost common set," in *Euro ITV*, Lisbonne, Portugal, Jun. 2011, p. irccyn contribution.

[89] Y. Pitrey, M. Barkowsky, P. Le Callet, and R. Pépion, "Evaluation of MPEG4-SVC for QoE protection in the context of transmission errors," in *SPIE Optical Engineering*, San Diego, United States, Aug. 2010.

[90] Y. Pitrey, U. Engelke, P. Le Callet, M. Barkowsky, and R. Pépion, "SUBJECTIVE QUALITY OF SVC-CODED VIDEOS WITH DIFFERENT ERROR-PATTERNS CONCEALED USING SPATIAL SCALABILITY," in *Third European Workshop on Visual Information Processing (EUVIP)*, Paris, France, Jul. 2011, p. paper number 67.

[91] M. Barkowsky, M. Pinson, R. Ppion, and P. L. Callet, "Analysis of freely available dataset for hdtv including coding and transmission distortions," in *Fifth International Workshop on Video Processing and Quality Metrics (VPQM)*, Scottsdale, United States, Jan 2010.

[92] P. Le Callet and E. Niebur, "Visual attention and applications in multimedia technologies," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2058–2067, Sept 2013.

[93] Z. Wang, L. Lu, and A. Bovik, "Foveation scalable video coding with automatic fixation selection," *Image Processing, IEEE Transactions on*, vol. 12, no. 2, pp. 243–254, Feb 2003.

[94] H. Sheikh and A. Bovik, "Image information and visual quality," *Image Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 430–444, Feb 2006.

[95] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297 – 312, 2011.

[96] S. S. Hemami and A. R. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design," *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 469 – 481, 2010, special Issue on Image and Video Quality Assessment.

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

[98] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, June 2008.

[99] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, Aug 2010, pp. 2366–2369.

[100] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From psnr to hybrid metrics," *Broadcasting, IEEE Transactions on*, vol. 54, no. 3, pp. 660–668, Sept 2008.

[101] U. Engelke, H. Kaprykowsky, H. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *Signal Processing Magazine, IEEE*, vol. 28, no. 6, pp. 50–59, Nov 2011.

[102] L. J. Karam, "Chapter 16 - lossless image compression," in *The Essential Guide to Image Processing (Second Edition)*, 2nd ed., A. Bovik, Ed. Boston: Academic Press, 2009, pp. 385 – 419.

[103] J. You, L. Xing, A. Perkis, and T. Ebrahimi, "Visual contrast sensitivity guided video quality assessment," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, July 2012, pp. 824–829.

[104] S. Winkler, "Analysis of public image and video databases for quality assessment," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 6, no. 6, pp. 616–625, Oct 2012.

[105] American National Standars Institute, "Objective video quality measurement using a peak-signal-to-noise-ratio (psnr) full reference technique," T1.TR.74-2001, American National Standards Institute, Ad Hoc Group on Video Quality Metrics, Washington, DC, Tech. Rep., 2001.

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

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

[108] P. Le Callet, C. Viard-Gaudin, and D. Barba, "A convolutional neural network approach for objective video quality assessment," *Neural Networks, IEEE Transactions on*, vol. 17, no. 5, pp. 1316–1327, Sept 2006.

[109] S. Tao, J. Apostolopoulos, and R. Guerin, "Real-time monitoring of video quality in ip networks," *Networking, IEEE/ACM Transactions on*, vol. 16, no. 5, pp. 1052–1065, Oct 2008.

[110] G. Van der Auwera, P. David, and M. Reisslein, "Traffic and quality characterization of single-layer video streams encoded with the h.264/mpeg-4 advanced video coding standard and scalable video coding extension," *Broadcasting, IEEE Transactions on*, vol. 54, no. 3, pp. 698–718, Sept 2008.

[111] K. Piamrat, C. Viho, J. Bonnin, and A. Ksentini, "Quality of experience measurements for video streaming over wireless networks," in *Information Technology: New Generations, 2009. ITNG '09. Sixth International Conference on*, April 2009, pp. 1184–1189.

[112] N. Staelens, S. Moens, W. Van den Broeck, I. Marien, B. Vermeulen, P. Lambert, R. Van de Walle, and P. Demeester, "Assessing quality of experience of iptv and video on demand services in real-life environments," *Broadcasting, IEEE Transactions on*, vol. 56, no. 4, pp. 458–466, Dec 2010.

[113] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 684–694, April 2013.

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

[115] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, Jan 2016.

[116] K. Zhu, C. Li, V. Asari, and D. Saupe, "No-reference video quality assessment based on artifact measurement and statistical analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 4, pp. 533–546, April 2015.

[117] F. Yang, S. Wan, Q. Xie, and H. R. Wu, "No-reference quality assessment for networked video via primary analysis of bit stream," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1544–1554, Nov 2010.

[118] J. Xu, P. Ye, Y. Liu, and D. Doermann, "No-reference video quality assessment via feature learning," in *2014 IEEE International Conference on Image Processing (ICIP)*, Oct 2014, pp. 491–495.

[119] Y. Li, L. M. Po, C. H. Cheung, X. Xu, L. Feng, F. Yuan, and K. W. Cheung, "No-reference video quality assessment with 3d shearlet transform and convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 6, pp. 1044–1057, June 2016.

[120] M. Dimitrievski and Z. Ivanovski, "No-reference quality assessment of highly compressed video sequences," in *Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on*, Sept 2013, pp. 266–271.

[121] F. Torkamani-Azar, H. Imani, and H. Fathollahian, "Video quality measurement based on 3-d. singular value decomposition," *Journal of Visual Communication and Image Representation*, vol. 27, pp. 1 – 6, 2015.

[122] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *2011 18th IEEE International Conference on Image Processing*, Sept 2011, pp. 2505–2508.

[123] R. Mok, E. Chan, and R. Chang, "Measuring the quality of experience of http video streaming," in *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on*, May 2011, pp. 485–492.

[124] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hobfeld, and P. Tran-Gia, "A survey on quality of experience of http adaptive streaming," *Communications Surveys Tutorials, IEEE*, vol. 17, no. 1, pp. 469–492, Firstquarter 2015.

[125] E. Baik, A. Pande, C. Stover, and P. Mohapatra, "Video acuity assessment in mobile devices," in *INFOCOM, 2015 Proceedings IEEE*, 2015.

[126] N. Damera-Venkata, T. Kite, W. Geisler, B. Evans, and A. Bovik, "Image quality assessment based on a degradation model," *Image Processing, IEEE Transactions on*, vol. 9, no. 4, pp. 636–650, Apr 2000.

[127] Z. Wang and A. Bovik, "A universal image quality index," *Signal Processing Letters, IEEE*, vol. 9, no. 3, pp. 81–84, March 2002.

[128] T. Brandao and M. Queluz, "No-reference quality assessment of h.264/avc encoded video," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 20, no. 11, pp. 1437–1447, Nov 2010, 54 citations.

[129] M. Narwaria, W. Lin, and A. Liu, "Low-complexity video quality assessment using temporal quality variations," *Multimedia, IEEE Transactions on*, vol. 14, no. 3, pp. 525–535, June 2012.

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

[131] M. Aabed and G. AlRegib, "No-reference perceptual quality assessment of streamed videos using optical flow features," in *Global Conference on Signal and Information Processing (GlobalSIP), 2014 IEEE*, Dec 2014.

[132] ——, "Reduced-reference perceptual quality assessment for video streaming," in *International Conference on Image Processing (ICIP), 2015 IEEE*, Sept 2015.

[133] ——, "PeQASO: Perceptual quality monitoring of streamed videos using optical flow features," *IEEE Transactions on Image Processing (submitted)*, 2016.

[134] H. A. Mallot, *Computational vision: information processing in perception and visual behaviour*. MIT Press, 2000.

[135] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–204, 1981.

[136] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'81. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1981, pp. 674–679.

[137] A. Bruhn, J. Weickert, and C. Schnörr, "Lucas/kanade meets horn/schunck: Combining local and global optic flow methods," *International Journal of Computer Vision*, vol. 61, no. 3, pp. 211–231, 2005.

[138] D. Fleet and Y. Weiss, "Optical flow estimation," in *Handbook of Mathematical Models in Computer Vision*, N. Paragios, Y. Chen, and O. Faugeras, Eds. Springer US, 2006, pp. 237–257.

[139] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Institute of Technology, May 2009.

[140] H. Ling and K. Okada, "Diffusion distance for histogram comparison," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, June 2006, pp. 246–253.

[141] M. Aabed and G. AlRegib, "No-reference quality assessment of hevc videos in loss-prone networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 2015–2019.

[142] M. Aabed, G. Kwon, and G. AlRegib, "POTUS: Perceptual video assement via power of tempospatially unified specra density," *IEEE Transactions on Circuits and Systems for Video Technology (submitted)*, 2016.

[143] ——, "Power of tempospatially unified spectral density for perceptual video quality assessment," in *submitted to the 2017 IEEE International Conference on Multimedia and Expo (ICME)*, Sept 2017.

[144] S. Wenger, "Nal unit loss software," in *Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG1, JCTVC-H0072*, feb 2012.

[145] Z. Long and G. AlRegib, "Saliency detection for videos using 3d fft local spectra," in *Proc. SPIE*, vol. 9394, 2015, pp. 93 941G–93 941G–6.

[146] "Structural similarity index (ssim) for measuring image quality - matlab documentation," http://www.mathworks.com/help/images/ref/ssim.html, accessed: 2015-11-25.

VITA

Mohammed A. Aabed obtained his PhD from the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA in 2016. He worked under the supervision of Prof. Ghassan AlRegib and conducted research in topics related to signal, images and video processing, multimedia and communications. His doctoral work is focused on video streaming in communication systems. Mainly, his doctoral research and thesis are concerned with the perceptual video quality assessment, especially in streaming application and the impact of streaming on the perceptual video quality under various coding conditions. He also conducted research on stereoscopic depth-based 3DTV where he worked on improving the streaming efficiency in DIBR-based 3D videos. He also worked on social signal processing and its applications in characterizing the behaviour and communication patterns of users in online social systems. He obtained his B.Sc. and M.Sc. in 2005 and 2008, respectively. His master's research was focused on call admission control and radio resources management in CDMA systems. During the summer of 2014, he worked at Samsung Research America as a research and development video coding engineer. In 2016, he also worked at Amazon Lab126 as a hardware development engineer for camera and vision systems. His research interests span image and video processing, coding and assessment, computer vision, relevant machine/deep learning and pattern recognition applications, multi-dimensional signal processing, compression and streaming, stochastic and probabilistic modeling of multimedia and communication systems and integrated services networks, social and communication networks analysis and modeling.

He is a member of the Multimedia and Sensors Lab (MSL) and Center for Signal

& Information Processing (CSIP) at Georgia Tech. He is also a member of the IEEE Signal Processing Society. He is the recipient of the CSIP 2016 Research Award.