Development of Video Quality Metrics Based on Psychovisual Models of Early Vision

A thesis

submitted in fulfilment

of the requirements for the Degree

of

Doctor of Philosophy in Engineering

at

The University of Waikato

by

Anastasia Mozhaeva

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2024

Abstract

Streaming video incurs many distortions during processing, compression, storage, and transmission, all of which can reduce the user's perceived video quality. Developing adaptive video transmission methods that increase the efficient use of existing bandwidth and reduce storage space while preserving visual quality requires quality metrics that accurately describe how people perceive distortion. A severe problem for developing new video quality metrics is limited data on how the human visual system processes spatial and temporal information simultaneously. The problem is exacerbated by the fact that the few data recognized by the scientific community, collected in the middle of the last century, used the ideas of obsolete display technology and were subject to medical intervention during collection, which does not guarantee a proper description of the conditions under which media content is currently consumed. As a result, modern video quality metrics do not provide stable and reliable data for predicting the subjective assessment of user quality. This research aims to investigate how the metrics being developed are made more efficient for assessing video quality by including new data from the psychophysical early vision model in the metrics.

The work proposed in this thesis comprises three main contributions: Firstly, the development of a novel method, software, and test equipment for research and measurement of the characteristics of the human visual system using modern display systems. Secondly, the refinement of the parameters o of a multidimensional model of human contrast sensitivity appropriate to modern display technology of viewing conditions. The contrast sensitivity function works like a filter through which visual stimuli must pass to be perceived by the observer. Only video artefacts in the passband region can be humanly perceived. Thirdly, the creation of a new full-reference and the first non-reference video quality metrics which consider the psychophysical features of the user's video experience. That provides stability in predicting the user's subjective rating of a video.

Among the three contributions of this thesis, first, a method for researching and measuring the characteristics of human visual systems on modern displays. In the proposed thesis, 27,840 visibility thresholds of spatio-temporal

sinusoidal variations were measured by a new method using different spatial sizes and temporal modulation rates. The obtained data is 96% more than any current contrast sensitivity function dataset and best describes a human's perception of video artefacts on a modern display. A multidimensional model of human contrast sensitivity in modern conditions of video content presentation is proposed for the first time based on new large-scale data and demonstrated that the presented visibility model has a distinct advantage for further development of media content transfer technologies. Since there is a limited number of video evaluation metrics based on fundamental knowledge about the work of the human visual system, a new full-reference metric is herein proposed. This proposed video quality metric extends the peak signal-to-noise ratio metric to include human visual system features. Finally, a new non-reference video quality metric that includes the psychophysical features of the user's video experience with stability in predicting the user's subjective rating of a video is proposed. The experimental results show that the proposed video quality metric achieves 81% more consistent performance in predicting user subjective quality among commonly used non-reference video quality metrics and comparable consistent performance to full-reference metrics on three independent video datasets.

The thesis also presents a large-scale database suitable for testing video streaming quality under video compression with artefacts, forming a learning base for future video quality metrics. The final dataset comprises 4.1 million video quality perceptual thresholds. The new database will contribute to the solution of a strong need for non-reference video quality metrics for user-generated video content to prevent loss of video quality caused by distortion during recording, compression and signal transmission.

Acknowledgements

First and foremost, I wish to express to my chief supervisor, Dr Lee Streeter, who has provided guidance and encouragement throughout the past three years. I am also grateful to my co-supervisor, Associate Professor Michael J. Cree, for the academic mentorship and support throughout the project.

I want to acknowledge my mentor, Associate Professor Igor Vlasuyk, who helped me grow over many years. I am also grateful to all my friends and colleagues. In particular, I would like to thank Sergei Romanov, Aleksei Potashnikov, Dmitrii Egorov, Vladimir Fedorov and Vladimir Mazin. I also want to thank Andrei Balobanov, who was, for me, an orientation in life values. Finally, I extend my wholehearted gratitude to my family for their encouragement that enabled me to succeed in this effort.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Video streaming occupies a growing share of internet bandwidth; in 2022, it amounted to 82% of Internet traffic [1]. An Internet-enabled HD television drawing a couple of hours of content per day generates as much internet traffic as the rest of the household today, on average. In 2023, 66% of connected flat-panel TV sets are 4K capable [1]. The characteristics of modern video display systems go beyond the capabilities of information transmission channels. Therefore, adaptive video compression methods that attempt to minimise visually noticeable losses are used today for modern ultra-high-resolution displays, virtual reality systems, etc. The main idea behind adaptive video streaming is to encode video content into multiple streams of various bitrate and quality levels that allow client-driven stream selection to meet the time-varying network bandwidth [2]. Depending on the available bandwidth, client equipment is adapted to different quality levels of video, and hence, users may be subjected to distortions [2]. Adaptive video transmission methods that optimise efficient use of available bandwidth and reduced storage space while preserving visual quality use video quality assessments as an integral part of quality correction in media streaming.

This thesis aims to increase the efficiency of the development of metrics for assessing video quality by including in the metrics the psychophysical model of early vision, namely, temporal and spatial characteristics at different back-

ground brightness levels. This will provide stability in predicting a user's subjective video rating.

1.1 Background

With the widespread adoption of digital television and Internet video streaming, video data compression processing to reduce the amount of information has become an essential component of broadcast and entertainment media [3]. However, lossy compression systems introduce compression artefacts into the final result, which leads to deterioration in the quality of the displayed video. Moreover, the limited bandwidth of physical channels causes some data to be deleted or skipped, also leading to deterioration in the quality of the resulting video.

Video quality metrics (VQM) are a core feature of modern video processing algorithms, are essential for optimizing processes in modern encoding systems, and have historically been the subject of much research. Despite this, the complexity of the human visual system does not allow for a simple algorithmic representation of a video quality metric [4]. Many modern applications remain poorly served by available video quality metrics, as they are based on often poor quality scientifically accepted data on the human visual system, which don't represent the actual human visual system (HVS) well. Moreover, these quality metrics do not consider certain aspects of modern displays, resulting in suboptimal performance [4].

The old-style analogue signal transmission used analogue electronic systems that are inevitably subject to noise generated by thermal and other sources. The system may also be subject to external sources of interference. As the signal amplitude decreases, the signal-to-noise ratio decreases [5]. The scheme for encoding and decoding video information during analogue video signal transmission is presented in Fig. 1.1. The analogue signal changed continuously. It was impossible to correlate algorithmically with human perception of the

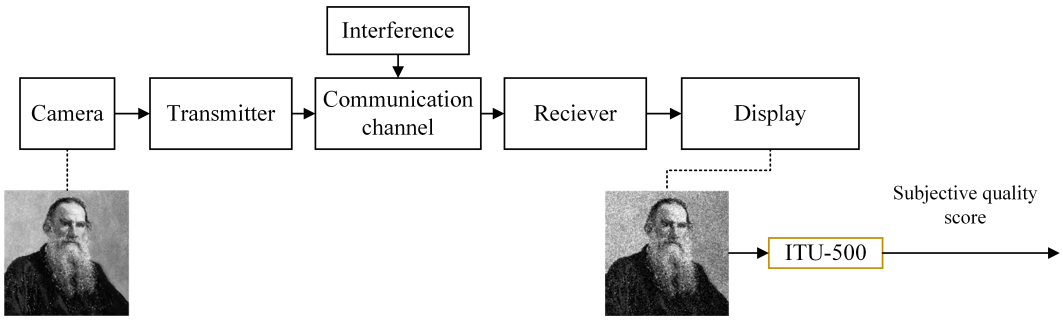


Figure 1.1: Scheme for encoding and decoding information during analogue transmission of a video signal using subjective user assessments according to the recommendations of the International Telecommunication Union (ITU).

transmitted frames in time and amplitude. Subjective user assessments were used according to the International Telecommunication Union (ITU) recommendations to determine the video's quality.

The introduction of digital broadcasting based on video encoded in MPEG-2 format has initiated a discussion on evaluating specific MPEG-2 artefacts for their impact on measuring video quality, where video quality is influenced by the parameters chosen for encoding. Transmission systems for digital video broadcasting had to exhibit a phenomenon known as error floor or quasi-error-free streams (QEF) of bits. Quasi-error-free (QEF) means less than one uncorrected error event per hour [6]. At this stage of the development of digital technologies, video quality assessments algorithmically predicted the correlation of specific MPEG-2 artefacts with parameters describing the subjective impression of the viewer [7], as shown in Fig. 1.2. The video quality metric, the peak signal-to-noise ratio (PSNR), was used to compare pixels in transmitted and received frames for consistency and has become widespread. The higher the PSNR, the better the quality of the compressed frame.

$$\mathrm{PSNR} = 10 \log_{10} \frac{(2^b - 1)^2}{\mathrm{MSE}} \tag{1.1}$$

where b is the number of bits per pixel value of the image. Mean squared error

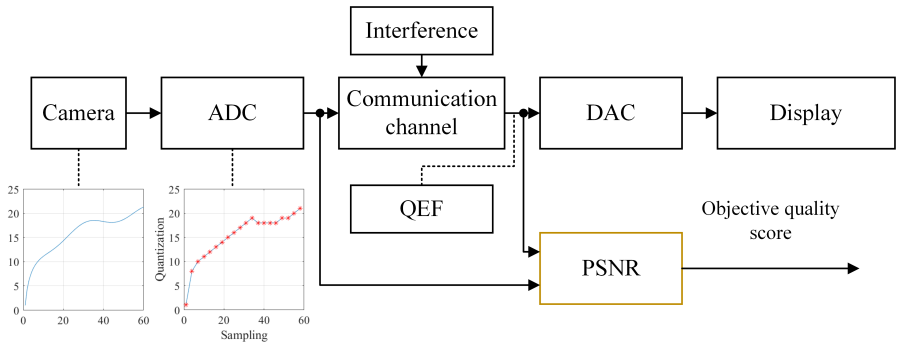


Figure 1.2: Scheme for encoding and decoding information digital broadcasting using PSNR to assess the effectiveness of eliminating interference during video signal transmission.

(MSE) determined by

$$\mathrm{MSE}(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{1.2}$$

where x $= \{x_i | i = 1, 2, , N\}$ and y $= \{y_i | i = 1, 2, , N\}$ are two finite-length, discrete signals (e.g., visual images), N is the number of pixels and x_i and y_i are the values of the i the samples in x and y, respectively.

PSNR measure is good because it is calculated by amplitude (in decibels) on a logarithmic scale. The eye perceives the signal on a logarithmic scale in amplitude. A two-fold increase in the signal amplitude does not mean the same improvement in frame quality for a human [8]. Today, PSNR is the most commonly used VQM in the world [9].

The next stage in developing video content transmission technologies included the following. First was the addition to the coding process of the psychovisual models of the human visual system. Second, the method of error control (FEC) during data transmission was included, in which the source (transmitter) sends redundant data, and the destination (receiver) recognizes only that part of the data that contains no visible errors [10].

At this stage of technology development, metrics for assessing video quality appear that imitate the work of the human visual system to evaluate the displayed video. The described information transmission method diagram is presented in Fig. 1.3. As seen from the scheme, encoders of this type use

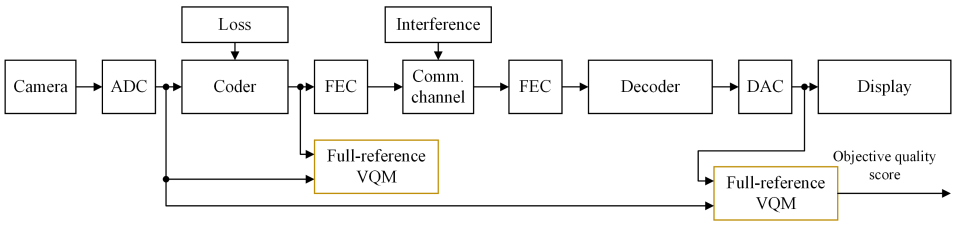


Figure 1.3: Scheme with two-pass encoding using video quality metrics which imitate the work of the human visual system.

a two-pass encoding scheme. Two-pass Encoding is a video file compression process where, in the first pass of encoding, the program analyzes the video file and creates a log file that contains information about the second pass. The second pass then uses this file to create an output video file with maximum quality. The second pass is slower as the program analyses the frame's quality and, if necessary, improves each frame. The analysis of the satisfactory quality of the frame occurs using the video quality assessment metric.

Today, video transmission protocols are based on network delivery protocols, where a user connects to a server, and the data can be transferred in either direction or both directions and flow control is provided. Video transmission uses acknowledgements and retransmissions to ensure that all bytes are received, no matter how long it takes [11]. Fig. 1.4 presents today's video transmission scheme. The figure shows that the decoder has information about the display device and the communication channel parameters. However, depending on the available bandwidth, client devices may adapt to different quality levels and therefore, users may suffer from compression/scaling artefacts and rebuffering when the available bandwidth drops. These video disruptions can negatively impact the user's Quality of Experience (QoE), which is the overall level of user satisfaction while viewing streaming content. Predicting QoE and acting on those predictions (determining the best combination of channel parameters and user satisfaction level) in encoding systems is important to improve the overall viewing experience. Understanding and predicting QoE, or in other words, non-reference video quality metrics for adaptive video streaming,

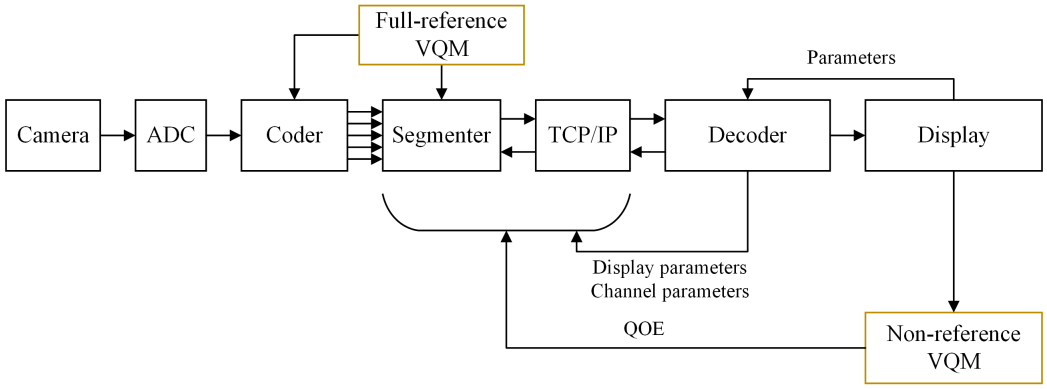


Figure 1.4: The scheme of video transmission protocols is based on the Transmission Control Protocol.

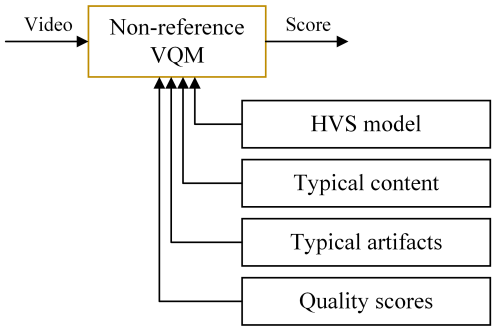


Figure 1.5: Components of a useful video quality metric in video transmission protocols today.

is an emerging area of research. Predicting video quality without full access to the reference video is a complex, unsolved problem with serious advantages for social networks and streaming media [2].

Today, to correctly form the quality levels in the encoder entering the segmenter, metrics must be used to assess video quality with linear stable prediction. This is necessary to determine the combination of parameters where the best bitrate stream gives the best QoE. For VQM to be useful in assessing video quality at a decoder, it must be explicitly related to the subjective perception of quality, see Fig. 1.5. Ideally, it should be standardized and independent of the systems or processes involved. Metrics necessary to assess video quality in the decoder during streaming must be processed independently in real-time [11].

1.2 Motivation and Objectives

Video quality metrics (VQM) are critical in modern streaming video processing algorithms to maximise bandwidth use and the user experience. In recent years, many methods have been introduced to assess video quality. Unfortunately, popular video quality metrics today correlate poorly with subjective perceptions of quality by the human visual system and depend on the systems or processes involved [3]. However, a small number of video quality assessments based on knowledge of aspects of the human visual system (affecting the perception of artefacts in the displayed signal) have achieved exceptionally high predictive rates [4, 9, 12, 13]. A critical issue is that when developing video quality metrics, it is necessary to represent the HVS aspects that define the artefacts perceived by the viewer in the software. However, the task is impeded by the limitations and contradictions of the old fundamental knowledge accepted by the scientific community about the perception of the displayed signal by a person and the lack of knowledge about the perception of HVS video content on modern equipment.

Another problem is that the VQM necessary for evaluating video quality during streaming should be independently processed in real-time and without human input when applied to a new video, or in other words, be a non-reference (NR) VQM [11]. The stable predicting video quality without full access to the reference video is a complex, unsolved problem with serious implications for social media and streaming media. To create (NR) VQMs, video datasets with large-scale human subjective ratings are needed to use machine learning efficiently. A wide range of video quality datasets have been released in recent years [14]. Nevertheless, current databases contain a small number of video sequences with little content diversity and distortion complexity, which offers limited support for developing and evaluating NR VQMs with efficient use of deep learning [15].

This thesis aims to answer the following research questions:

- Is it possible to increase the efficiency of metrics for assessing video quality by including in the metrics the psychophysical model of early vision, namely, temporal and spatial characteristics at different background brightness levels?

- Is it possible to provide stability in predicting a user's subjective video rating in streaming video?

To achieve this goal, the following scientific and practical tasks are presented in this work:

- Analysis of the architecture of modern video codecs to identify characteristic problems in the encoded video signal;

- Development of equipment for measuring and analyzing the characteristics of human visual systems on modern displays based on research on the fundamental limitations of human vision;

- Research and development of a new full-reference video quality metric of video quality based on the characteristics of the human visual system;

- The development of a methodology and equipment for measuring the quality of encoded video using an acceptable minimum threshold of the user's perception, which enables the generation of video sequences of constant quality;

- Research and development of a new non-reference video quality metric that includes the psychophysical user perception of temporal and spatial characteristics at different background brightness levels, thereby providing stability in predicting the user's subjective video rating.

This work uses modern methods of digital video processing, statistical radio engineering, spectral analysis and programming to solve the assigned problems.

1.3 Principal Contributions

This thesis makes the following original contributions:

- A new concept for creating metrics for evaluating video quality, taking into account the peculiarities of the human visual system when watching video on modern display devices. The new concept builds upon previously existing video quality ratings from the literature. It addresses the limitations of video quality metrics, such as the unstable positive correlation of algorithmically predicted video quality with subjective human perception of video quality.

- A new method, software and test equipment for the study and large-scale measurement of the characteristics of the human visual system that affects the perception of artefacts by a person when viewing video content. The method measures spatio-temporal sinusoidal variations using different spatial dimensions and temporal modulation rates at different brightness levels. The method considers modern display devices' features and allows the collection of large-scale data that was not obtainable before.

- Refinement of the parameters of a multidimensional model of human contrast sensitivity in modern conditions for providing media content, based on 27840 user perception thresholds. New knowledge can significantly accelerate progress in video compression and quality, video rendering systems, and video masking.

- A new full-reference video quality metric extending the most famous peak signal-to-noise ratio metric by considering the characteristics of the human visual system as measured using modern screens. A simple and easily repeatable metric comparable to the best complex algorithmic metrics.

- A new device for measuring encoded video quality using an acceptable minimum perceptual threshold allows the generation of video sequences

of constant quality. The new approach allows for obtaining well-labelled data in a larger volume in a short time compared to the current methodologies.

- A new constant-quality video dataset created using the dynamically changing visual quality of typical media content contributes to addressing the shortage of well-labelled video datasets.

- An illuminance-adjusted spatio-temporal video quality score accurately determines the local spatio-temporal video quality. The proposed video quality metric achieves the highest results of prediction of subjective perception for different video contents among commonly used non-reference metrics and comparable prediction with full-reference metrics [16].

1.4 Thesis Overview

This work is divided into six chapters and one appendix, with the conclusion and future outlook located in Chapter 7.

The development theory and implementation of modern video quality assessments are reviewed in **Chapter 2**. The limitations of the transmission of video content today are presented. The fundamentals of objective and subjective quality assessment and psychometric scaling are covered. The current state of the art of video quality metrics is reviewed. The causes of poor correlation of algorithmically predicted video quality with subjective human perception of video quality are discussed. The reasons for the limited data on how the human visual system perceives artefacts in video, processing spatial and temporal information, are revealed.

In **Chapter 3**, the system design process of measurement of the characteristics of human visual systems that affect the perception of artefacts by a person when viewing video content. Existing measurement methodologies represent a limited amount of data. These reasons are behind the motivation to build a new method, software and test equipment for the large-scale

measurement of the characteristics that take into consideration the features of modern display devices. Contrast thresholds were measured at different spatial and temporal frequencies at different background brightness levels. The experimental results are analysed and discussed.

In **Chapter 4**, the refinement of the multidimensional model of human contrast sensitivity parameters for modern video transmission is presented since there are a limited number of video evaluation metrics based on fundamental knowledge about the operation of the human visual system, a new full-reference metric was also created to test the model, expanding the metric of the peak signal-to-noise ratio by the characteristics of the human visual system. The results of testing the proposed model and existing visibility models are presented and analysed.

In **Chapter 5**, a new device for measuring encoded video quality by users using an acceptable minimum perceptual threshold is developed. A solution for obtaining well-labelled video data with subjective scores in a larger volume in a short time is developed. A new constant-quality video dataset has been generated using the dynamically changing visual quality of typical media content during subjective tests.

Using the proposed visibility model and a new set of constant quality video data to improve and stabilize the performance of non-reference video quality measures is shown in **Chapter 6**. An illuminance-adjusted spatio-temporal non-reference video quality metric has been created and tested. The proposed approach provides the most consistent high performance among widely used non-reference metrics and comparable performance stability to full-reference metrics.

1.5 Publications

The following publications resulted from this PhD:

- **A. Mozhaeva**, L. Streeter, I. Vlasuyk, A. Potashnikov,“ Full Reference Video Quality Assessment Metric on Base Human Visual System Consistent with PSNR,” 28th Conference of Open Innovations Association (FRUCT), Russia, 2021, pp. 309–315.

- **A. Mozhaeva**, I. Vlasuyk, A. Potashnikov, M. J. Cree and L. Streeter, “ The Method and Devices for Research the Parameters of the Human Visual System to Video Quality Assessment,” 2021 Systems of Signals Generating and Processing in the Field of on Board Communications, Russia, 2021, pp. 1–5.

- **A. Mozhaeva**, A. Potashnikov, I. Vlasuyk, L. Streeter,“ Constant Subjective Quality Database: The Research and Device of Generating Video Sequences of Constant Quality,” 2021 International Conference on Engineering Management of Communication and Technology (EMCTECH), Austria, 2021, pp. 1–5.

- **A. Mozhaeva**, I. Vlasuyk, A. Potashnikov, L. Streeter,“ Full reference objective metric for assessing video quality compatible with PSNR considering the frequency and peripheral characteristics of human vision”, Digital Signal Processing Application Considerations, Russia, vol. 2, 2021, pp. 44–54.

- **A. Mozhaeva**, V. Mazin, M. J. Cree, L. Streeter,“ Video Quality Assessment Considering the Features of the Human Visual System.” Image and Vision Computing (IVCNZ). Lecture Notes in Computer Science, New Zealand, vol. 13836, 2022.

- **A. Mozhaeva**,“ Video Quality Assessment Adapted For TV Signals Considering Modern Media Content Transmission Features,” Radio engineering, *In press.*

- V. Mazin, M. J. Cree, L. Streeter, K. Nezhivleva, **A. Mozhaeva**,“ Research and Application of the Adaptive Model of the Human Visual Sys-

tem for Improving the Effectiveness of Objective Video Quality Metrics," 33rd Conference of Open Innovations Association (FRUCT), Slovakia, 2023, pp. 192–197.

- **A. Mozhaeva**, V. Mazin, M. J. Cree, L. Streeter, " NRspttemVQA: Real-Time Video Quality Assessment Based on the User's Visual Perception", 38th International Conference on Image and Vision Computing New Zealand (IVCNZ), New Zealand, 2023, pp. 1-7.

- **A. Mozhaeva**, M. J. Cree, I. Vlasuyk, A. Potashnikov, L. Streeter," A contrast sensitivity model of the human visual system in modern conditions for presenting video content," Plos One, vol. 19(5): e0303987, 2024.

The following publications were published during the period of the PhD study but are not part of the work presented in the thesis:

- A. Potashnikov*, **A. Mozhaeva***," Research of the threshold resolution of the human visual system for the primary colours of the RGB system," DSPA: Issues of digital signal processing, vol. 11, no. 1, 2021, pp. 3–10. *(*) Authors had an equal contribution.*

- A. Potashnikov, **A. Mozhaeva**, I. Vlasuyk, V. Fedorov, A. Balobanov," The Method of Forming a Panorama with Increased Resolution in the Spatial Direction," Systems of Signals Generating and Processing in the Field of on Board Communications, Russia, 2021, pp. 1–5.

- A. Potashnikov, V. Mazin, N. Stepanov, A. Smirnov, **A. Mozhaeva**," Analysis of Modern Methods Used to Assess the Quality of Video Sequences During Signal Streaming," Systems of Signals Generating and Processing in the Field of on Board Communications, Russia, 2022, pp. 1–4.

- A. Drebuzhan, K. Nezhivleva, A. Potashnikov, I. Vlasuyk, **A. Mozhaeva**, " The Steganographic Method of Introducing Additional Information,

Resistant to Transcoding and Conversion," Systems of Signals Generating and Processing in the Field of on Board Communications, Russia, 2022, pp. 1–5.

- D. Egorov, V. Fedorov, I. Vlasuyk, A. Potashnikov, **A. Mozhaeva**," A Novel Method for Estimating the Spatial Frequency Characteristics of Cameras Based on Generative Random Sequences," Systems of Signals Generating and Processing in the Field of on Board Communications, Russia, 2022, pp. 1–5.

- K. Nezhivleva, A. Davydova, A. Drebuzhan, **A. Mozhaeva**, A. Balobanov, " Comparing of Modern Methods Used to Assess the Quality of Video Sequences During Signal Streaming with and Without Human Perception," Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO), Russia, 2022.

- M. Vyatkin, A. Potashnikov, V. Selivanov, I. Vlasuyk, K. Nezhivleva, **A. Mozhaeva**, " Method of preventing leakage of personal data through eye-tracking modules of user devices, T-Comm, Russia, vol. 16, no. 7, 2022, pp. 44–51.

- A. Egorova, R. Baryshev, **A. Mozhaeva**," Methodology of Researching Perception Identity of Regions of Users' Interests While Viewing Streaming Video Containing Various Content and Compression Artefacts," Systems of Signals Generating and Processing in the Field of on Board Communications, Russia, 2023, pp. 1–7.

Chapter 2

Video Quality Metrics

At the present state of technology development, video coding systems exhibit high-quality and completely satisfactory results. However, video streaming continues to capture an increasing share of Internet bandwidth. With video traffic skyrocketing, improving video encoding technology is critical for video streaming companies because the parameters of contemporary video display systems go beyond the capabilities of information transmission channels. To solve the problem today, adaptive video transmission methods are used. When developing strategies for adaptive video transmission, video quality estimates are integral to correcting the quality of adaptive video compression systems in multimedia streaming.

In this chapter, compression methods are first considered, and the limitations of video content transmission today are presented. It also describes the principles of operation and the theoretical foundations of objective and subjective quality assessment. The current state of video quality metrics is considered, and the reasons for the poor correlation of algorithmically predicted video quality with the subjective perception of human video quality are investigated.

The work in this Chapter was presented in [16, 17, 18] during this doctoral research project.

2.1 Video Compression

The digital video represents a natural (real) visual scene filmed in space and time. Video encoding is the process of compressing and decompressing a digital video signal [19]. In general, compression methods can be divided into two main classes: lossless compression and lossy compression. Lossless compression is a reversible method that allows you to recover the exact original information from compressed data, for example, as used in medical or legal images. However, lossless methods produce larger files than lossy methods, limiting their utility. Lossy methods provide for greater compression by reducing the amount of information transmitted and are generally based on the fact that the human visual system is less sensitive to high-frequency distortion and colour differences than low-frequency distortion and luminance differences. Information not perceived by the human visual system is ideally removed or reduced. However, inaccuracy in the process affects the quality of the transmitted data.

2.1.1 Basic Video Codec Model

The video encoder's operation principle, see Fig. 2.1, consists of three main functional blocks: a temporal model, a spatial model, and an entropy encoder [19]. Here, we explain each block in turn.

Temporal model. The input to the temporal model is an uncompressed video sequence. The temporal model aims to reduce redundancy between transmitted frames. It happens by forming a predicted frame and subtracting the predicted frame from the current frame [19]. Using the previous frame as the predictor for the current frame is the simplest method of temporal prediction. However, the problem is that much energy remains in the residual frame. In practice, the motion compensation method is widely used, which compensates for the movement of rectangular areas of the current frame. Block-based motion compensation is relatively straightforward and computationally tractable.

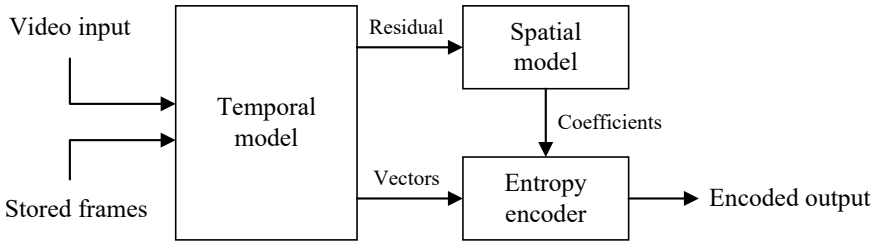


Figure 2.1: Video encoder block diagram.

In some cases, a better prediction of motion compensation can be constructed based on interpolated (intermediate) positions of samples on a reference frame.

Spatial model. The spatial model uses similarities between neighbouring samples in the residual frame to reduce spatial redundancy. The quantized transform coefficients are the output of the spatial model [19].

Entropy encoder. The entropy encoder is used for compressed parameters of the temporal model (motion vectors) and the spatial model (coefficients). However, natural objects rarely have sharp rectangular boundaries, and many kinds of motion are difficult to compensate for using block methods. Even in encoders of efficient algorithms, there is a problem of incorrectly finding motion vectors [19].

Data reconstruction. In addition to encoding and transmitting each piece of the frame, the encoder reconstructs it to provide a reference for further predictions [19].

2.1.2 Current Video Compression Technologies

H.264/MPEG-4. The first standards, developed by the MPEG group, such as Motion-JPEG and MPEG-1, are still widely used for storing compressed video files on PCs and the Internet, see Fig. 2.2. The first working standard for videotelephony, H.261 and H.263, was created by the Video Coding Experts Group (VCEG). The MPEG group united its efforts with the VCEG to de-

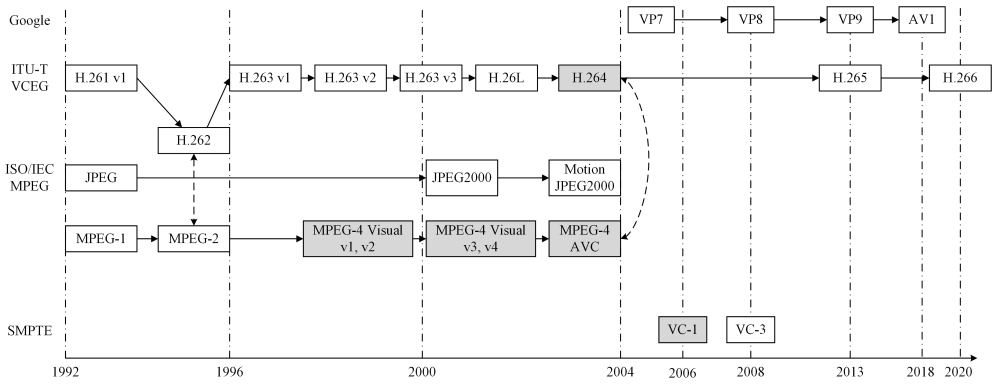


Figure 2.2: Standardization of multimedia compression formats.

velop the H.264/MPEG-4 standard. The H.264 and MPEG-4 standards share several common features and are the most commonly used now in the world. H.264/MPEG-4 uses block-based motion compensation transform, quantization, and entropy coding [19]. The general scheme for eliminating spatial statistical dependence for MPEG-4 Visual and H.264 standards defines the compression process of a single video stream frame and coincides with the procedure for compressing a static full-colour image using the JPEG standard. MPEG-4 and H.264 are open international standards; any person or organization can purchase the documents from the International Organization for Standards. At the same time, there are tens of thousands of patents related to video coding, and it is impossible to implement standards without potentially infringing patent rights [19].

High-Efficiency Video Coding (HEVC). HEVC (H.265) was developed as part of a joint video project between VCEG and MPEG in 2013 to solve two critical problems of H.264: increased video resolution and increased use of parallel processing architectures. The new standard has received features such as coding block quadtree structure, adaptive sample bias, advanced motion vector prediction, etc. [20]. In H.265, Coding Tree Units can use a large block structure of up to 64 $\times$ 64 pixels. A block can be continuously divided into coding units using the coding tree unit quadtree syntax. In this way, H.265 can adapt to high-resolution video encoding. In inter-frame prediction, advanced

motion vector prediction is used. H.265/HEVC doubles the compression ratio compared to H.264 with the same visual perception of quality level [20].

Standard H.266. 6 Jul 2020 Fraunhofer HHI is "proud to present the new state-of-the-art in global video coding: H.266/ Versatile Video Coding" [21]. The primary objective of Versatile Video Coding is to improve compression performance over the existing HEVC standard significantly. Based on the HEVC standard, Versatile Video Coding refined existing technologies and added novel coding tools. Versatile Video Coding has a block-based hybrid coding architecture as in most preceding standards.While HEVC only supports blocking using four trees, Versatile Video Coding uses more complex and varied partitioning modes. The quadtree has been replaced with a quadtree plus binary tree structure [22]. Quadtree plus binary tree extends the four-part division by optionally splitting into a binary tree in the vertical or horizontal direction to provide more flexible video partitioning. A multi-type tree structure has four splitting types: vertical binary splitting, horizontal binary splitting, vertical ternary splitting, and horizontal ternary splitting. A coding unit can have either a square or rectangular shape in the coding tree structure, see Fig. 2.3. Quadtree plus binary tree introduces a binary separation that divides the current coding unit into two sub-coding units of the same size, horizontally or vertically, along the middle [23]. A generalized binary partition extends this concept by allowing you to move the separation boundary perpendicular to the separation direction, see Fig. 2.4. Predicted samples of luma components are obtained using four-tap filter interpolation. In contrast, HEVC uses linear (2-tap) filtering. H.266 has extended merge prediction, merge mode with motion vector differences.Versatile Video Coding uses a transform coding scheme with a special form of trellis quantization, a low-complexity version of vector quantization, and transform coefficient coding with four passes in a subblock. Besides the deblocking filter and sample adaptive offset (the two-loop filters in HEVC), an adaptive loop filter with block-based filter adaption is applied. In general, the encoding system at this stage of technology development is

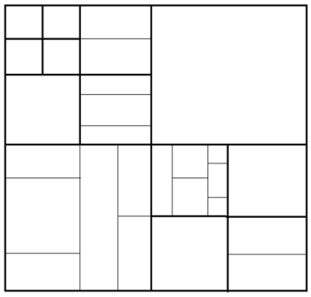


Figure 2.3: Examples of partitioning in versatile video coding.

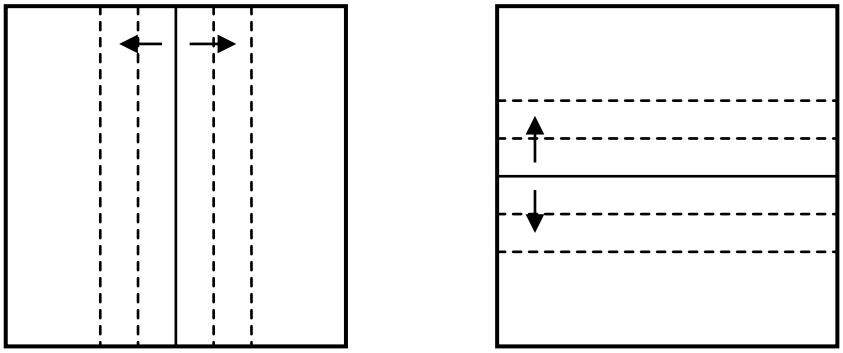


Figure 2.4: Versatile video coding test model: quadtree plus binary tree.

quite good, and as can be seen from the description of the standards presented above, H.266 will improve the basic components of previous versions.

AV1. Google and YouTube widely use the benefits of Google VP9's codec to improve its video services. VP9 considerably outperforms H.264 and is also a strong competitor to H.265. The AV1 video compression format was developed at the end of 2018. AV1 includes high-level syntax and parallelization features. VP9 uses a four-way partition (64 $\times$ 64 level down to the 4 $\times$ 4 level) [24]. VP9 codec uses a tree-based boolean non-adaptive binary arithmetic encoder to encode all syntax elements. H.265 and H.264 require licensing fees, and open-source solutions are almost impossible for commercial products. The amount of royalties is even more significant for companies engaged in providing services for video on demand and systems with many videos for storage and transmission, such as Google from YouTube, Netflix, and so on. All this creates a demand for free codecs, such as Google VP9, designed to provide competitive patent-unencumbered compression efficiency. It is also worth noting that,

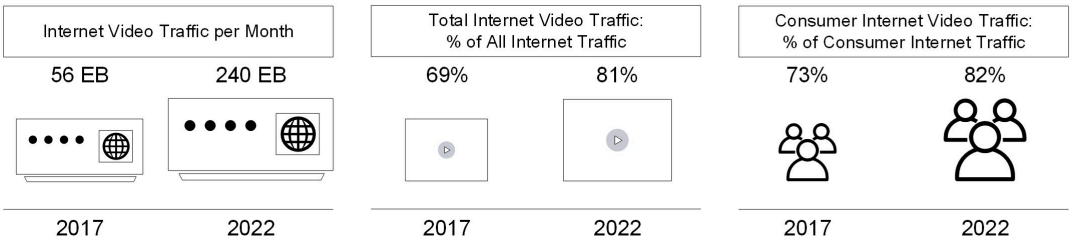


Figure 2.5: Change in video content consumption by users 2017-2022.

despite its complexity, modern video compression is based on simple psychovisual models that imitate the work of the human visual system by analogy with JPEG.

Neural network compression. Neural network compression has become a popular topic today. The main difference from the techniques presented above is that the original image is converted to feature space and converted back to the reconstructed image [25]. The differences between the original and reconstructed images are used as the loss function for training neural networks. Developing research in the field of video coding based on neural networks is focused in two different directions: (1) improving existing video codecs by making better predictions that are included in the same codec structure and (2) holistic end-to-end image/video compression schemes. The holistic methods proposed for video compression are very different from each other and use substantially different approaches. The only way to compare new methods based on neural networks is to look at the results. However, none of the currently proposed methods has resulted in a breakthrough that would justify replacing the existing, well-studied video compression framework. Given the state of research in this area, it is too early to determine the most promising direction or to compare the proposed algorithms according to generally accepted criteria [25].

2.1.3 The Critical Problems of Video Compression Methods

At this stage of development, compression systems have a thoroughly studied structure and generally good performance. However, in 2021 video traffic from content delivery networks occupied 71% of all bandwidth consumed, and in 2022, it already accounted for 82% of Internet traffic [1], Fig. 2.5. Internet-connected HDTV broadcasting a couple of hours of content daily generates, on average, as much Internet traffic as the entire household today. In 2023, 66% of connected flat panel TVs support 4K [26]. Today, the world faces a problem where the parameters of modern video display systems go beyond the capabilities of information transmission channels even when using modern advanced coding methods. Therefore, the adaptive video transmission system is widely used today for modern ultra-high resolution displays, virtual reality systems, etc.

However, even in adaptive video transmission systems, the problem of video compression and transmission artefacts still remains open. For example, the most popular codec is H.264, which provides a lower level of noise suppression than the less common MPEG-4 Visual. It should also be considered that the main spatial distortions are block distortions and blurring of the image within the processing unit. A characteristic temporary distortion is the flickering effect, noticeable to the observer at a sufficiently low encoding rate. In addition to block distortion, it is necessary to note the Gibbs effect, which is noticeable near the borders and contours of the frame. Coarse quantization of high-frequency conversion coefficients also results in wave-like noise visible when panning slowly in a highly detailed scene, such as a crowd patch at a sporting event. In wavelet encoders, at very low encoding rates, an interfering temporal frequency becomes noticeable due to the accumulation of errors in a group of frames. Spatial distortion, characteristic of all wavelet codecs, is the blurring of image details when encoding at low speed.

The main idea of adaptive video streaming is to encode video content into multiple streams with different bitrates and quality levels and allow stream selection based on the client according to time-varying network bandwidth. Therefore, Internet transmission capacity and storage capacity may be reduced. However, choosing the best combination of channel parameters and user satisfaction levels remains an open question. For example, a video quality metric with unreliable predictions may underestimate user satisfaction at high bitrates and, conversely, overestimate satisfaction at low bitrates. The client device, responsible for deciding the quality level of the video segment that is to be played next, needs accurate knowledge about the perception of artefacts by the human visual system under various communication channel parameters.

Consequently, an integral part of adaptive video compression systems is the video quality assessment to correct streaming media by network bandwidth. Several definitions of the quality of the displayed information include Jacobson's description [27],“ the subjective impression formed in the mind of the observer of the degree of excellence exhibited by an image”. The implication is that the quality of the video cannot be separated from the opinion of the viewer, that it is fundamentally subjective, and that it is the result of a combination of several different perceptual attributes. It indicates the complexity inherent in determining video quality. It is not easy to describe something that is ultimately a subjective impression in the observer's mind [27]. To optimize the transmitted video quality and create an effective video quality score, one must first understand how viewers perceive distortion caused by compression and transmission.

Since video quality assessment is an integral part of adaptive video compression systems today, it is necessary to understand how modern video quality assessments work and their limitations.

2.2 Video Quality Assessment

Numerous quality assessment (QA) methods have been proposed over the past years, which can be generally classified as subjective and objective quality assessments. Subjective quality assessment is carried out by a person and regulated by several international standards [28, 29, 30]. For example, ITU-R BT.500-15 proposes the methodology of collecting a subjective quality assessment for television signals. Objective quality assessment is an evaluation performed algorithmically. They are discussed further in the following.

2.2.1 Subjective Quality Assessment

Since the human is the end user in most digital video processing applications, subjective visual analysis is the most relevant but of lesser accuracy quality assessment method [31]. The area of subjective evaluation concerns the measurement and modelling of human judgments. The perceptual processes are analyzed by studying the experience or behaviour of the subject due to systematic changes in the properties of the stimulus along one or more physical dimensions [32]. In quality assessment experiments, participants evaluate a set of stimuli or conditions according to some criteria.

Subjective quality assessment methods [29, 28] can be generally classified as rating and ranking methods (or comparative judgment), Fig. 2.6. Rating methods allow participants to rate incentives using a categorical or continuous interval scale. The rating scale is usually not universal and may require additional calibration to adjust the ratings received from individual observers [33]. The quality can be expressed in rating experiments using the following model [34]. The rating π_{ik} for participant k and condition i. That is, it depends on the scalar q_i, a measure of the truth quality; the random variable δ_k is the bias of the participant; and a random variable ξ_{ik}, the inaccuracy of the participant and the complexity of stimulus estimation.

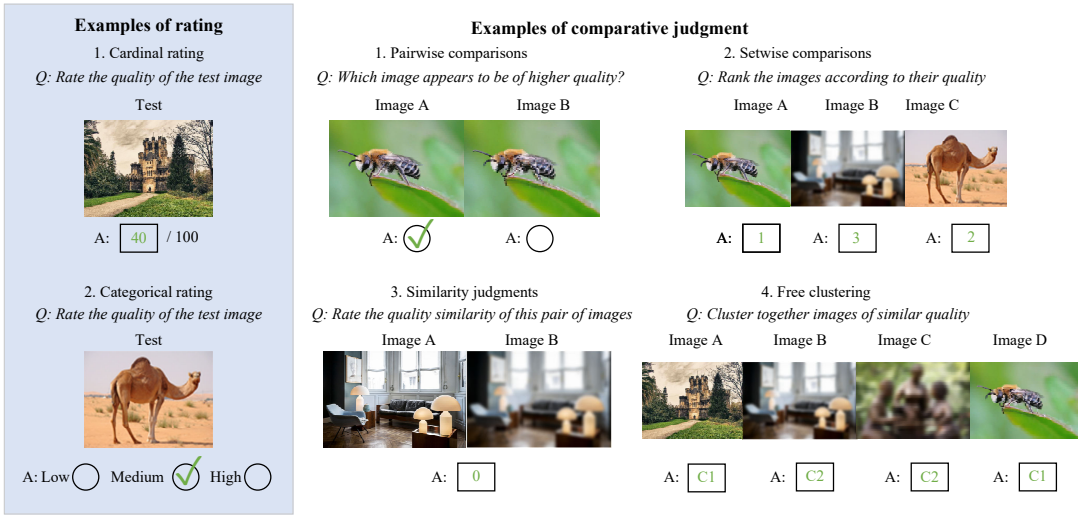


Figure 2.6: Examples of different subjective judgment experiments.

$$\pi_{ik} = q_i + \delta_k + \xi_{ik} \tag{2.1}$$

The bias and imprecision components in the model are assumed to be independent random variables that are normally distributed, and ξ_{ik} is assumed to have zero mean.

Pairwise comparison experiments have less effect on cognitive load and require little preparation. However, the total number of possible comparisons increases quadratically with the number of stimuli. Thurstone [32] and Bradley-Terry [35] models are often used for pairwise comparisons, providing similar solutions. For example, in Thurston's models, the perceived quality i is modelled as a random variable ω_i:

$$\omega_i \sim N(q_i, \beta_i^2) \tag{2.2}$$

where the mean of the distribution is the true measure of quality q_i, and the standard deviation β_i considers the combined variance between participants.

Obtaining a scale based on rating experiments does not require an optimization procedure and is based on averaging the ratings of individual observers for each condition. This average is called the Mean Opinion Score (MOS). The result of subjective expert assessments is influenced by the nature of the

displayed information and various external factors. It is assumed that the quality rating may vary depending on which image is familiar to the evaluator. On unfamiliar images, compression and transmission distortion may not be noticeable. However, for a familiar image, the estimates will be more illustrative because the observer knows the structure of the image. For the quality assessment results to be reproducible, the International Telecommunication Union adopted the Regulatory Document ITU-R BT.500-15 [28], which contains instructions for conducting subjective tests, test materials, and rating scales.

Although accurate, subjective studies are seriously flawed as a stand-alone measure of video quality because

- Subjective scores are time-consuming and expensive. Subjective results can only be obtained in the course of experiments with a large number of participants.

- Subjective scores cannot be used in real-time applications such as media compression and transmission systems.

- The assessment results largely depend on the participant's physical and emotional states. Moreover, other factors, such as the display device and lighting conditions, also affect the results of such experiments.

2.2.2 Objective Quality Assessment

Objective quality assessments aim to develop mathematical models that predict the user's perceived video quality. An effective VQA maximises bandwidth usage and user experience while avoiding costly user research. VQA's researchers attempt to match computer predictions generated by algorithms with subjective video viewing ratings obtained from users through experiments.

Objective video quality metrics can be divided into full-reference (FR-VQM) when all information about the undistorted reference frame is available;

reduced reference, when partial information is available (RR-VQM); and non-reference, when there is no information about the reference frame (NR-VQM).

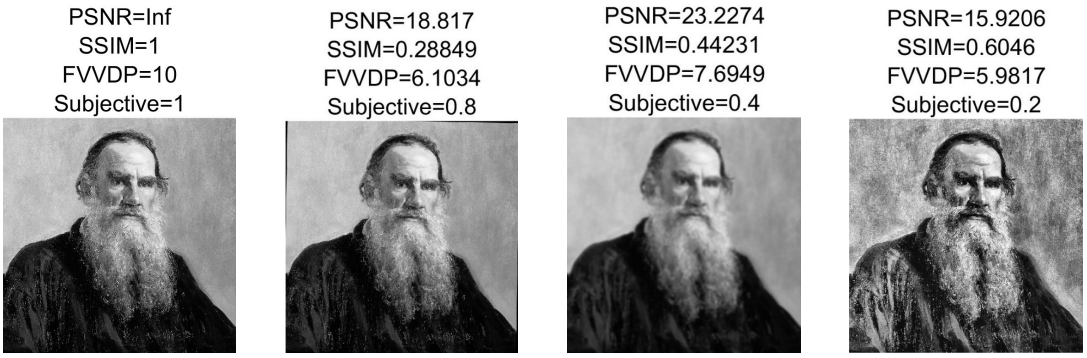
Full-reference image and video quality metrics for quality prediction:

- Methods based on the structure of the frame [36, 37, 38, 39, 40, 41, 42, 43, 44]. Structural similarity methods are top-down full-reference metrics that aim to quantify distortion without knowing anything about the characteristics of the human visual system (HVS) [11]. According to Z. Wang:“ The most fundamental principle underlying structural approaches to image quality assessment is that the HVS is highly adapted to extract structural information from the visual scene, and therefore, a measurement of structural similarity (or distortion) should provide a good approximation to perceptual image quality” [45].

The most popular quality assessment based on the structure of the frame is the Structural Similarity Index Measure (SSIM) [45]. The formula calculates the value of the structural similarity measure:

$$SSIM = \left(\frac{\sigma_{xy}}{\sigma_x \sigma_y}\right)\left(\frac{2\bar{X}\bar{Y}}{(\bar{X})^2 + (\bar{Y})^2}\right)\left(\frac{2\sigma_x\sigma_y}{\sigma_x^2\sigma_y^2}\right) \tag{2.3}$$

where SSIM is the value of the measure of similarity (quality) of images; $X = \{x_i j\}$ and $Y = \{y_i j\}$ are compared images; M, N - image sizes. In general, a measure of image similarity is calculated in disjoint areas for each image separately. Since the objects in the scene images did not change or move, the measure can be calculated immediately on the entire image. However, for fragments of large or small values of medium brightness, local SSIM estimates are not stable. This is especially noticeable when images are compressed using the JPEG2000 algorithm. SSIM do not take account of different absolute luminance levels or viewing distance. SSIM cannot accurately assess image quality, but only the similarity of the two images (do not correlate well with human perceptions of image quality) and does not always correctly assess the similarity of the depicted scenes. Over the last few years, numerous variations of SSIM and other algorithms that estimated quality based on structural



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

Figure 2.7: Comparison of video quality metrics for Lev Tolstoy's frame altered with different distortions. a) Reference frame. b) Reference frame with one-degree counterclockwise rotation with bicubic smoothing. c) Reference frame with Gaussian blur distortion. d) Reference frame with adaptive histogram equalization with contrast limiting.

similarity.

- Methods based on other structure measures [46, 47, 48, 49, 50, 51, 52]. Several objective quality algorithms, such as gradient methods, are used [11].

- Methods based on statistics [53, 54, 55, 56, 57, 58, 59]. Quality measures have been proposed that are based primarily on statistical/information-theoretic indicators, often supplemented by machine learning methods [11].

The most popular FR-VQMs are peak signal-to-noise ratio (PSNR) and SSIM [45]. These full-reference objective VQMs are easy to understand and implement, but they have the significant drawback of poor correlation with the human visual system, see Fig. 2.7. FR-VQMs based on aspects of the HVS have proven superior [4, 60, 61].

In reduced-reference video quality metrics, several features are extracted from a reference frame and used as additional information for the evaluation. An essential parameter in the development of an RR-VQM is the data rate. If a high data rate is available, it may include more information about the reference frame, allowing more accurate quality predictions to be made. On

the other hand, if a low data rate is used, then only a small amount of reference image information can be transmitted. RR-VQM should provide an efficient generalization of the reference frame, be sensitive to various types of distortions, and have good perceptual relevance.

The vast majority of algorithms of Non-reference video quality metrics try to detect certain kinds of distortion, such as blurring, blocking, ringing, or various forms of noise:

- Methods for Blurriness/Sharpness [62, 63, 64, 65].

- Methods of certain compression artefacts [66, 67, 68, 69, 70, 71, 72]. For example, methods for JPEG compression artefacts. The general approach of such methods involves measuring the edge strength at block boundaries. The quality is then determined based on this perceived blockiness score [11].

- Methods, general purpose, usually use classifications and regressions, where the regressors/classifiers are trained using certain features. The scores obtained from subjective experiments are helpful in training and testing NR-VQMs. Data-driven metrics, such as those based on CNNs [73, 74, 75, 76, 77, 78, 79] are built on automatic feature extraction. However, complex machine learning models depend on the quality and quantity of training data. If there is little data, the model cannot generalize. To alleviate the problem of insufficient and noisy data, transfer learning is often used, which can improve the ability of metrics to generalize. However, this does not replace the need for larger and more diverse datasets.

For objective VQA to be useful in assessing video quality, it should be demonstrably related to subjective perceptions of quality. It should ideally be standardized and independent of the systems or processes involved. The VQA necessary for evaluating video quality during streaming should be independently processed in real-time, or in other words, be a non-reference VQA [11]. Stable predicting video quality without full access to the reference video is a complex, unsolved problem with serious implications for social media and streaming media. Unfortunately, popular NR-VQA prediction models do not

work well with distorted real-world videos [73].

A significant part of the existing image and video quality metrics characteristics are presented here.

2.2.3 Video evaluation research path

Over the past decade, the focus of video quality research has shifted from the broad goal of understanding how people evaluate video quality to the more limited purpose of computer design algorithms simulating human subjective assessments obtained in experiments [11]. Based on these studies, many methods for evaluating video quality have been created and implemented, from free algorithms to commercial products, the latter of which are increasingly used in industry for various applications.

The ITU-T Group on Audiovisual Quality Assessment (IRG-AVQA) aims to coordinate progress on specific topics of interest, limited to video and audiovisual image quality assessment, both subjective and objective [80, 81]. The Video Quality Expert Group (VQEG) was founded in 1997 by the ITU-T and ITU-R groups. VQEG is advancing the field of video quality assessment by evaluating objective quality measures and exploring new subjective evaluation methods. One of the areas of VQEG's work is video quality assessment based on human psychophysiology, where new psychophysiological methods and methodologies are being created to assess video quality in real-time. To date, VQEG has conducted several areas of research:

- Computational video quality models based on psychophysiological measurements.
- Accurate quantitative relationships between physiological signals and visual quality.
- Research on optimal methodology (extensive literature research in measuring psychovisual thresholds).
- Creation of a representative data set.

However, at the moment, only limited data on the relationship between physiological signals and perceived information quality is available, namely, data on psychovisual thresholds for a person's perception of the displayed information. Also, new methodologies for collecting representative datasets are still needed.

However, the Laboratory for Image and Video Engineering (Live) published the following review about VQEG: "In our view, the video quality assessment tools advocated by VQEG and occasionally blessed by the ANSI, International Organization for Standards, and/or ITU standards bodies based on internal VQEG tests have never caught hold in commercial practice (including VQM) since they are not held to adequate public scrutiny and scientific repeatability" (information is presented on the official website of Live [82]). LIVE explores information-theoretic approaches to the problem of video quality assessment. The laboratory indicates that using many known psychophysical features of the human visual system is possible, but the currently available HVS data is still difficult to verify. At Life, research focuses on developing improved objective video quality models that accurately match subjective quality scores. Some of the lab's work is to work on VQM in a limited area of high frame rate video and to determine whether it is possible to combine quality scores from multiple time ranges using the temporal contrast sensitivity function.

The designers of quality assessment algorithms and video compression must decide how psychophysical outcomes relate to quality in modern conditions (i.e., on modern display technology) for presenting video content without completely understanding how presented video content is perceived by the human eye. Creating large-scale datasets of the human visual system acceptable for video quality designers remains an unresolved problem. New knowledge and tests on the relationship between the HVS and VQA are needed [11].

Over the past decade, the focus of video quality research has shifted from the broad goal of understanding how people evaluate video quality to the more limited goal of developing computer algorithms that imitate the human subjec-

tive scores obtained in experiments [11]. In recent years, many methods have been introduced to assess video quality. However, an exceptionally high degree of prediction was achieved by a small number of video quality assessments (VQA) based on knowledge about aspects of the human visual system (HVS) that affect the perception of artefacts of the displayed signal [4, 13, 51, 60]. Today, to create a highly predictive video quality assessment, it is necessary to understand which characteristics of the human visual system need to be used and their limitations.

2.3 Modelling visual perception

The basics of human vision concerning psychology include fixation of information with the eye or early vision; perception, where object recognition and mental manipulation of information occur; the cognitive part of seeing, or, in other words, the awareness of what is seen. Integrating knowledge of early vision into computer-aided design tools can significantly improve the experience of multimedia content for the average web user [83].

Temporal dynamics, spatial processing and colour processing, characterise early vision. The information processing of the human visual system in early vision can be described using three segments: filtering, encoding and interpretation. Filtering determines information collected and lost throughout the system or on a specific stream or channel [83]. Encoding describes how specific visual mechanisms represent specific components of visual information. Interpretation describes how encoded information, possibly from multiple sources, including memory, is used to determine the state of objects in the visible world. Modern research has paid enough attention to the coding and interpretation segments and much less filtering [83].

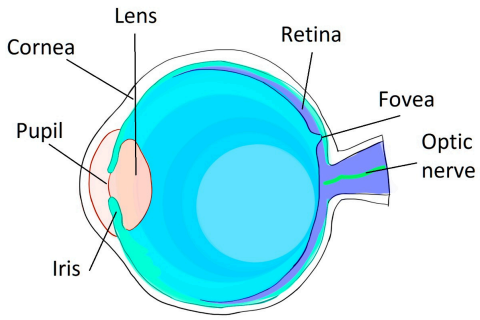


Figure 2.8: Diagram of a horizontal cross-section of the human eye.

2.3.1 Structure of the human visual system

Human visual perception, in the form to which we are accustomed, is possible due to the anatomical structure of the human eye. Fig. 2.8 presents a schematic representation of the structure of the human eye, showing the light-sensing elements - the retina, the fovea, the optic nerve, and the eyeball the focusing part of the eye with the cornea, the lens, the iris, the pupil [84].

The optical image formed by the eye is projected onto the retina. The retina is a part of the brain that is separated from it in the early stages of development but is still connected to it through a bundle of fibres – the optic nerve. A layer of cells on the back surface of the retina serves to convert information into chemical and electrical signals. These signals are then processed and transmitted to the brain via the optic nerve. This layer of cells contains light-sensitive receptors known as rods and cones.

The functioning of the eye can be described as the focusing of light on the retina of the human eye to form an image that is detected and transmitted to the brain while, in turn, interpreting the received optic nerve impulses to extract useful information. Interestingly, the optic nerve consists of approximately one million fibres, carrying information generated by approximately 130 million elements. The result of the joint work of the eye and brain is the visual scene as we perceive it, with all the complexity of shapes, depth, movement, colour and texture.

Rods, significantly more numerous (about 120 million per retina) than

cones (about 7 million per retina), are responsible for our vision in low light and are saturated in daylight. Cones do not respond to low light but are responsible for the ability to see fine details and for colour vision, also called daylight vision. The number of rods and cones varies markedly in different parts of the retina. There are only cones in the centre, where our vision's ability to discern fine detail is greatest. This rodless area, about half a millimetre in diameter, is called the central fovea.

2.3.2 Luminance adaptation

The stimulus for vision is light distributed in space and time. The distribution of each of these parameters influences our visual experience. In the most general form, in the physics of forming a perceived stimulus relative to the human visual system, everything that affects our visual sensations and reactions can be considered. This may be a state of adaptation to light, perception of the boundaries of objects in the field of view, distance from the viewed object, or flickering.

Luminance is the amount of light emitted or reflected toward the eye per unit area of the source, weighted by the photopixel's luminous efficiency function, expressed in candelas per square meter. It has been experimentally proven that the luminance perceived by humans is a logarithmic function of the physical luminance of the light entering the eye. The human visual system can adapt to a huge range of illumination [85].

Another important aspect of perception is chromatic adaptation, which adapts the visual system to colour. Chromatic adaptation can be considered analogous to automatic white balance in video cameras.

2.3.3 Contrast Detection

The fundamental limits of the HVS were first described by Luizov almost 60 years ago [86]. These limits were found through experimental and theoretical studies of the contrast threshold: the smallest contrast reliably perceivable

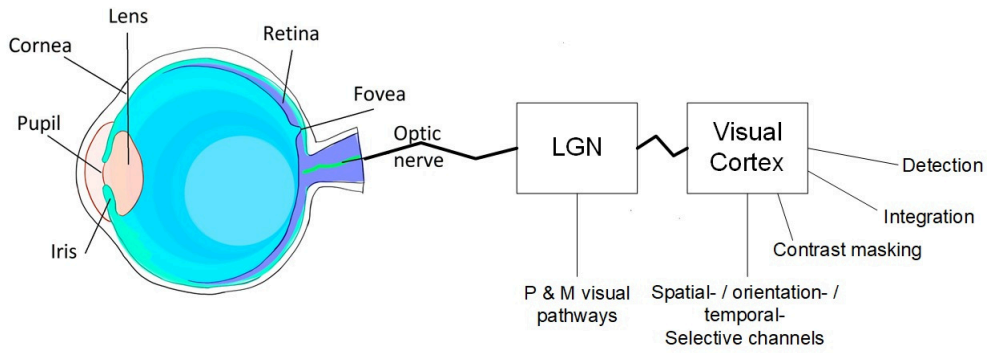


Figure 2.9: Contrast sensitivity function.

under given conditions. Contrast sensitivity is defined as the inverse of the contrast threshold [87].

The contrast sensitivity function (CSF) can be considered a bandpass filter [88], Fig. 2.9. A plot of contrast sensitivity as a function of frequency is called a contrast sensitivity function, Fig. 2.10. In spatial frequency, it is the spatial contrast sensitivity function (SCSF). In temporal frequency, this is the temporal contrast sensitivity function (TCSF) [87]. The CSF characterises the earliest stage of vision that occurs 100 to 120 ms after the presentation of the stimulus to the observer's eye [89]. Early vision includes capturing, preprocessing, and coding visual information and excludes interpretation or other cognitive processing of visual information [83]. (More recent definitions of early vision, for example, that given by Cecchi (2018), include computation of basic properties like shape and colour.) Because the CSF is a bandpass filter through which visual stimuli must pass to be perceived by the observer, only video artefacts in the passband region can be humanly perceived. Hence the importance of CSF to this study.

The HVS remains extremely sensitive to contrast throughout most of the visual perception. Human contrast sensitivity studies have a special place in experimental psychology studies. However, as noted above, any relationship to video quality is not discussed in detail in such research.

Sade [90] first proposed the method for measuring the eye's reaction to sinusoidal gratings. The work models the human visual system as an analogue

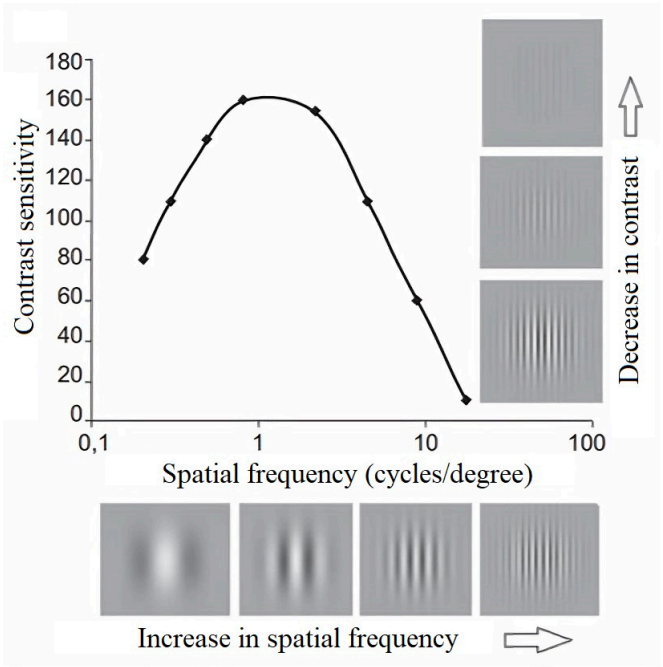


Figure 2.10: Spatial contrast sensitivity function.

sequential system with a single channel bound by the limitations of the optical elements of the eye and the post-retinal systems [91]. Lange [92] measured the temporal contrast sensitivity function in the range of adaptation of the retinal illumination for a single spatial target of indefinite duration. Volkov [93] measured the threshold of spatial contrast sensitivity using black and white gratings with a smooth sinusoidal profile of brightness changes. The sinusoidal shape of the grating is recognized as preferable, despite the complexity of its manufacture, since, firstly, according to the Fourier theorem, any wave complex, including a rectangular one, can be transformed into a series of sinusoidal waves. Secondly, any defocusing that reduces the contrast of such waves does not violate their shape. In his experiments, Volkov varies the thickness of the stripes, which determines their spatial frequency, expressed by the number of black-and-white cycles per degree. In 1962, Fergus Campbell began his highly regarded contrast sensitivity studies. As stimuli in his experiments, sinusoidal gratings of different contrast and spatial frequency were used, presented on the screen of a cathode-ray tube. Threshold contrasts were measured by changing the grating's modulation depth. As a result, Campbell and Robson [94]

obtained a frequency-contrast characteristic.

Unfortunately, since the focus of previous research has not been directed at video compression problems, research covering all areas of interest to assess video quality has not been carried out. Extant datasets on CSF have been collected as temporal measurement threshold only [95]; threshold of spatial contrast sensitivity only [96, 93]; and simultaneous spatial and temporal measurements using targets that were sinusoidal in both space and time, but with independent spatial and temporal frequencies [94]. The problem is that the visibility of temporal signals is probably not separable from their spatial configuration, so purely temporal measurements are of limited practical use [83]. Several efforts have been made to combine both spatial and temporal models of visibility [97, 98, 87, 4, 91]. However, it must be considered that the data were obtained in independent experiments for TCSF and SCSF, which does not guarantee the accuracy of the model construction.

2.3.4 Spatiao-Temporal CSF Models

Kelly proposed a spatio-temporal CSF [97], in which the system responds to moving stimuli that differ in spatial and temporal frequency. To study the joint sensitivity to time and space, experiments were performed to assess the observer's ability to detect and judge the direction and estimate the velocity of motion in sinusoidal gratings. The temporal frequency of the moving sinusoidal grating was the number of sinusoidal cycles of luminance variation per second that can be registered at any given point. Kelly's data were collected for stimuli that strongly shifted contrast sensitivity towards higher frequencies. Daly adjusted the latter [98] to describe naturally better-observed stimuli (used CRT display). However, Kelly-Daly's model does not account for any other aspects of the CSF [91].

Watson and Ahumada presented a linear model of space-temporal contrast sensitivity of the HVS, which they termed "the pyramid of visibility" [87], Fig. 2.11. The pyramid of visibility describes the spatial and temporal con-

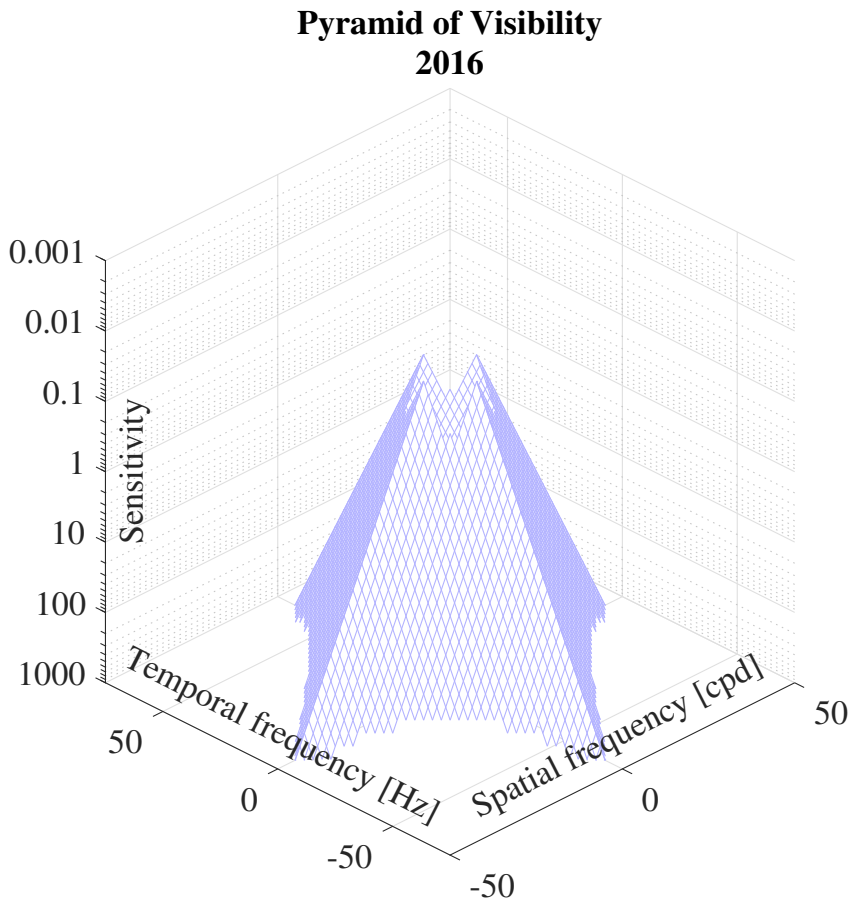


Figure 2.11: The pyramid of visibility [87].

trast sensitivity and its dependence on retinal illumination, all based on the modest results of much older studies [99, 95, 94, 100]. According to Watson and Ahumada: "Human contrast sensitivity is a linear function of spatial frequency, temporal frequency, and the log of adapting retinal illuminance at photopic retinal illuminances at moderate to high frequencies. A linear model means that the variation in sensitivity with one of the variables, temporal frequency, spatial frequency, or log retinal illuminance, is independent of the other two. The linear formulation describes the complete surface as the "pyramid of visibility." The height of the pyramid rises linearly with the log of retinal illuminance. As a result, the window of visibility is always a diamond that grows and shrinks linearly, with the log of retinal illuminance" [87]. However, it was noted by the authors that a more comprehensive survey of the relevant parameter space, both from the literature and from new data, is

Table 2.1: The modern datasets on the operation of the CSF HVS.

Dataset	Datapoints
Modelfest [Watson 2000] [103]	14
HDR-VDP CSF [Mantiuk et al. 2011] [104]	86
HDR CSF [Wuerger et al. 2020] [105]	204
Rovamo et al. [1993] [106]	46
Robson [1966] [107]	97
Laird et al. [2006] [102]	8
Snowden et al. [1995] [108]	222
Virsu and Rovamo [1979] [109]	31
Virsu et al. [1982] [110]	72
Wright and Johnston [1983] [111]	209
Anderson et al. [1991] [112]	89

desirable [87]. Also, the model is limited in not covering the full spatial and temporal frequency range perceived by the human visual system.

In 2021, the Computer Laboratory of the University of Cambridge, together with Facebook Reality Labs, presented the FovVideoVDP model [4], which is based on a combination of the spatial-chromatic CSF [101], the cortical magnification model and Kelly-Daly's model [102]. At the testing time, the video quality metric (VQM) with the FovVideoVDP model achieved the best correlation with human video quality assessment [91]. However, later, it was indicated that the FovVideoVDP model incurs a prediction error for low and high spatial frequencies.

In 2022, the unified stelaCSF model was presented that considers all the main stimulus parameters, including spatial and temporal frequency and luminance. The model was generated by combining data from the early eleven measurements [91], presented in Table 2.1. The use of older measurements of the CSF is due to the difficulty of conducting new large-scale experiments of

the human visual system.

2.3.5 Limitations

All the above information and the complexity of existing experiments provide a paucity of available HVS data to describe how people perceive distortion on modern displays. New knowledge and tests of the relationship between HVS and the perception of compression and information transfer artefacts are needed [11]. These data should control for the spatiotemporal frequency and brightness in one experiment [43].

The use of older measurements of the CSF (in recently proposed VQMs) is due to the difficulty of conducting new large-scale experiments of the human visual system. To measure CSF in three dimensions for 10 points, in each measurement, where each measurement takes 90 seconds (which is typical for 30–50 trials), leads to $10 \times 3 \times 90/3600 = 25$ hours per experiment for one observer [91]. When using an optimized Bayesian adaptive method from medicine and psychophysics for studying thresholds of human perception of spatial and temporal change [113], it takes about 30 evaluations for a typical detection problem, or two hours to find 20 thresholds [114, 115]. However, suppose it is necessary to find ten mutual combinations between three parameters of brightness, spatial and temporal contrast thresholds, being an acceptable amount of data for using machine learning. In that case, the Bayesian adaptive method requires 30,000 tests, amounting to 1 year of testing for one evaluator! The Bayesian adaptive method does not scale up to fine-grained multi-feature data collection; therefore, creating a relatively comprehensive and sufficiently accurate model of the CSF faces the practical obstacle of large-scale and extremely time-consuming experiments, as summarised in Table 2.1.

It should also be noted that the viewing conditions in older studies significantly differ from the displays in common use today. First, the field of view was very small, which is uncharacteristic of modern video presentations. Second, a device based on a cathode ray tube was typically used to obtain the

stimulus (CRT). The main disadvantages of CRTs are well-known [116]:

- low static contrast (a CRT screen is usually quite bright in the absence of an image),

- low spatial resolution,

- dependence of the point scattering function on brightness,

- low stability of CRT parameters over time (due to changes in cathode emission and phosphor luminosity, uneven brightness and contrast across the field).

It should be noted that, to a certain extent, the shortcomings are compensated for in high-precision display devices, but usually to obtain the best user experience rather than improve parameters and characteristics.

Probably the most significant disadvantage of CRT stimulus formation is the pulsed nature of image formation, in which the electron beam scans the screen and is intensity-modulated, thus transmitting the brightness of each point to the screen. The screen, depending on the intensity of the electron beam, begins to fluoresce. The increase in fluorescence intensity with time when the screen is bombarded is not long, depending on the intensity of this impact and its duration (which, however, is approximately constant or corrected to such because the curvature of the screen does not correspond to the radius to it from the electronic projector) on the intensity. The decrease in the luminescence intensity concerning the time of the phosphor after the cessation of irradiation is considered to be exponential. Therefore, to avoid temporal distortion of the stimulus due to a long afterglow, it is necessary to use a phosphor with a short afterglow time. Consequently, in experiments, the formation of a stimulus in time and space occurs due to the temporal integration of stimulus structure in the human visual system using the temporal characteristic, which is simultaneously proposed to be studied. The stimulus is a short pulse of light with a short increase in intensity and a longer, but also short, decrease, shifted in space-time along the CRT screen (in fact, by the phase characteristic) by the scanning law used. To assess the sensation

of brightness of such a stimulus, Talbot's law usually uses a simplified model of the characteristic being studied [117]. In this case, possible nonlinearities in the perception of a stimulus and effects associated with the duration of the afterglow are usually not taken into account, even though, for example, prolonged perception of an image on a CRT monitor screen causes observer fatigue to a greater extent than, for example, a similar static image illuminated source of constant lighting over time. Estimating the degree of measurement error in stimulus synthesis by the method used when using a different display technology is beyond the scope of this work.

The advantage of a monochrome CRT is the absence of a pixel structure, the formation of a stimulus not by a triad of primary colours but by a phosphor colour predetermined during production. However, as mentioned above, the most significant disadvantage of the CRT-based setup used in the past is its relatively long time of subjective experiment.

In addition, in most psychophysical experiments, the person is withdrawn from the natural environment to stabilize the tests or, in other words, to reduce the number of experiments and, consequently, errors. For example, Kulikowski used eye drops during the experiment [99], which changed the actual perception of the participants.

2.4 Summary

In this chapter, the limitations of video content transmission, despite the impressive advances in the field of compression and, were discussed, as was the need for adaptive video transmission (2.1 Video Compression). Developing adaptive video transmission methods that provide increased bandwidth and reduced storage space while preserving visual quality requires video quality metrics that accurately describe how people perceive distortion (2.2 Video Quality Assessment).

Features of the human visual system are revealed that affect the correlation

of algorithmically predicted video quality with a person's subjective perception of video quality (2.3 Modelling visual perception). A severe problem for developing new video quality metrics is the limited data on how the early human visual system simultaneously processes spatial and temporal information (2.3.5 Limitations).

Today, with the development of statistical apparatus and data science to create accurate models of CSF, it is desirable to have a large amount of initial data for processing since the model's simplicity and design are no longer significant factors determining its popularity in practical use. These are the main data on which my thesis is based.

New experimental measuring equipment of the human visual system parameters that affect the correlation of algorithmically predicted video quality with a person's subjective perception of video quality will be created in Chapter 3. In Chapter 4, an improved CSF model will be presented that overcomes all the limitations indicated in this chapter, such as an independent collection of data on spatial and temporal frequency, not a lot of data, and irrational use of equipment from older studies today. Using the proposed visibility model and the new set of constant quality video data created in Chapter 5 will allow multimedia researchers to get closer to solving the problem of non-reference video quality indicators identified in this chapter. Also, a new spatio-temporal light-corrected video quality metric will provide the most consistently high performance among commonly used non-reference metrics and comparable consistent performance to fully reference metrics.

Chapter 3

Experimental Methodology to Measure the Human Contrast Sensitivity Function

The model of contrast sensitivity of the human visual system has applications in

- Video compression systems. The parameters of modern video display systems go beyond the capabilities of information transmission channels, so today, video compression methods are used without visually noticeable losses based on knowledge of human visual perception. This applies to modern ultra-high resolution displays, virtual reality systems, etc. [91].
- Video quality assessment systems. Methods for assessing video quality are becoming integral to adaptive video compression systems and adaptive correction of video quality in media streaming [4].
- Various systems in which it is crucial to attract human attention, including industrial display systems or, for example, creative techniques used in the visual arts and cinema.
- Systems and devices for visual masking of objects, applicable, for example, in architecture.

As shown in Chapter 2, the focus of previous research has not been directed at video display problems, where knowledge about the perception of vision by users on modern display systems is required. That is, there is a lack of studies covering all areas of interest to developers of video compression and evaluation methods. Today, developers of video quality assessments and video compression algorithms need to determine how psychophysical results relate to quality in today's video content presentation environment without a full understanding of the processes of the human visual system. Extant datasets on contrast sensitivity function (a bandpass filter through which visual stimuli must pass to be perceived by the observer; only video artefacts in the passband region can be humanly perceived) have limitations. Based on research on the limit to human vision, in this chapter, we present a new method, software, and test equipment for researching and measuring the characteristics of the human visual system (contrast sensitivity function).

This chapter presents methodology by forming stimuli with a specified spatial and temporal characteristic depending on brightness and evaluating the response HVS to created stimuli. Using a more efficient approach avoids the complexity of the implementation of the experiment and receives a large amount of data. Each participant only requires about 30 minutes to perform 960 evaluations of various interactions of luminance, spatial, and temporal frequency values and temporal frequency values.

The work in this Chapter was presented in [60, 17] during this doctoral research project.

3.1 Stimulus and Apparatus

We employed a modern liquid crystal matrix display using In-Plane-Switching (IPS) technology and computer control [17], thereby eliminating the disadvantage of CRT technology used in older studies. The limitations of CRT-based installations given in Chapter 2 and the fact that such devices have long been

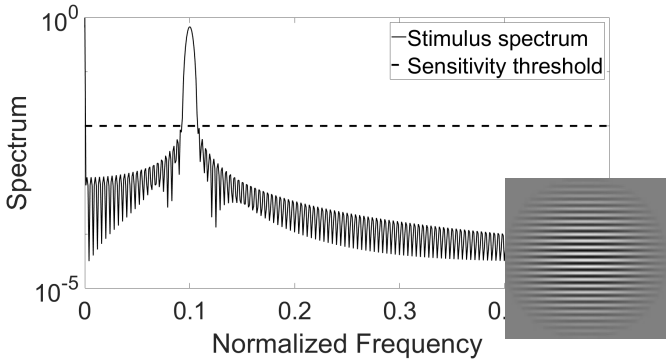


Figure 3.1: The stimulus used to generate Mira. Normalized frequency, is the frequency of the spectrum. The horizontal dashed line represents the spatial sensitivity threshold.

discontinued mean that experiments today would require a custom-made CRT with high spatial resolution and, at the same time, a high sweep rate. Based on a balanced decision about the shortcomings and advantages of hardware today, it was decided to manufacture a display device based on a serial display. The IPS technology is chosen because it allows the introduction of temporary changes in information by modulating the brightness of the backlight, which is possible only based on liquid crystal technology. IPS matrices have better contrast and colour accuracy within wide viewing angles than CRT. The disadvantages usually include a longer response time, but in our case of demonstrating static stimuli, this is an advantage on a dynamic screen.

Most previous data sets of contrast sensitivity function were determined from experiments with cosine (or sine) gratings modulated in the spatial and temporal dimensions [91]. A grating's size is typically limited by a square or circular aperture, which a Gaussian function was usually used to restrict. The non-uniform test pattern is the round sinusoidal lattice shown in Fig. 3.1, dubbed Mira. The test stimulus is round within the field of a centre vision so that all parameters of the test stimulus are symmetrical [17]. The disk stimulus and the experimental approach are consistent with previous psychovisual approaches to measuring contrast sensitivity function [91]. During the experiment, the spatial characteristics of vision need to be changed in each

direction, while all parameters of the test stimulus, except for the spatial frequency, should be independent of the direction in which the measurements are taking place. Therefore, the test stimulus can only be round.

The goal of contrast detection experiments is to find the smallest contrast at which the stimulus is detected, depending on the experimental procedure. In this experiment, a method of adjustment (psychometric protocol) is used, in which the observer directly controls the contrast and confirms his choice as soon as the stimulus becomes visible or invisible [91]. The subjective experiment is to visually compare the test object with the background and find the minimum contrast at which the horizontal lines of Mira become distinguishable. The large number of thresholds obtained levels out of high variance caused by subjective judgments. The CSF is designed to predict only low-contrast, barely visible (near-threshold) differences shown on a uniform background. The experiments use the limit method, in which the stimulus is initiated at zero (undetectable) modulation and is then gradually increased in intensity by the participant until the perceptual threshold of the stimulus is found. Finding the threshold is comparable to quality assessment metrics, where video content providers use solutions that only allow videos of acceptable quality to the user [2]. Since the measurement is carried out by the threshold method, the threshold sensitivity of the visual system is about 1% of the maximum brightness value [119]. In preliminary testing of equipment, it was found that the monitor's spatial resolution was sufficient to measure the thresholds, namely, to visually compare the test object with the background and further find the minimum contrast for normal observers.

The development software interface allows for the adjustment of the monitor parameters (pixel brightness, amplitude, and period values for both temporal and spatial parameters) and the selection of the image type for analysing and measuring the characteristics of the HVS. Fig. 3.2 shows a general view of the interface. The brightness, flicker amplitude, and flicker period values can be set in the temporal parameters field to study temporal information. This

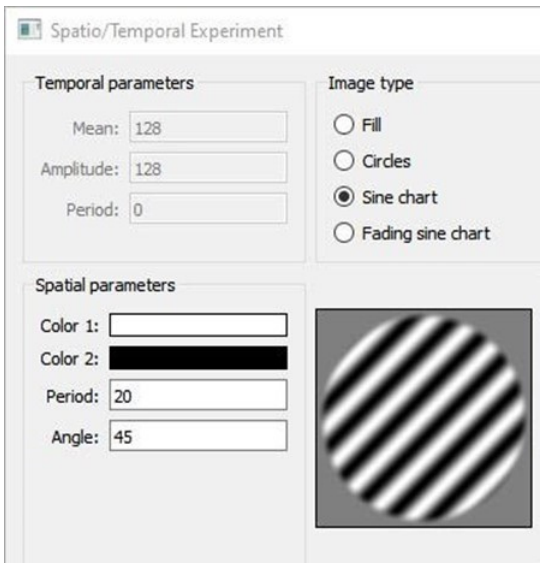


Figure 3.2: Software interface. The monitor parameters are pixel brightness, amplitude, and period values for both temporal and spatial parameters.

block is used for experiments related to the perception of monitor flicker. To study spatial information, the contrast values (color1, color2) and period of the Mira can be set in the spatial parameters field. Spatial parameters are used to analyze the HVS spatial threshold sensitivity.

This interface has three images: Fill, Circles, Sine, and Sine attenuation. An experiment is started after the selection of the experiment type from the Editable Options drop-down list below the window where the image is displayed. The flicker amplitude position is intended to measure temporal information, and the spatial amplitude position is intended to measure spatial information. This list allows changing the value of the required parameter using the arrows on the keyboard (right and left arrows). The software and test equipment are designed to study the temporal characteristics of the human visual system based on the frequency dependence of the flicker amplitude for various brightness values. The program also included a block for the dependence of spatial frequency on amplitude for black-and-white images, which measures contrast depending on spatial parameters.

Let w and h be the pixel width and height of the display with centred coordinates x_0 and y_0 given by,

$$x_0 = x - \frac{w}{2}, \qquad \text{and} \qquad y_0 = y - \frac{h}{2}. \tag{3.1}$$

Let d be the diameter of the test pattern in pixels, then r, the normalized centre-to-point distance, since during the experiments, the test signal must fall into the central vision area, is given by

$$r = \frac{1}{d}\sqrt{x_0^2 + y_0^2}. \tag{3.2}$$

To assess the spatial resolution, the normalized pixel brightness of the test stimulus, where the background brightness is 0.5:

$$M(x_0, y_0) = 0.5 \sin \frac{2\pi t(x_0, y_0)}{T} + 0.5. \tag{3.3}$$

where t defines the phase of the spatial oscillation. The window function localises the test signal's energy in the main lobe. The Kaiser window is used here:

$$w_0(x) \triangleq \begin{cases} \frac{1}{L}\frac{I_0\left[\beta\sqrt{1-(\frac{2x}{L})^2}\right]}{I_0[\beta]} & |x| \leq \frac{L}{2} \\ 0 & |x| > \frac{L}{2} \end{cases} \tag{3.4}$$

where β is a Kaiser window parameter [118], L is the window duration, and I_0 is the modified Bessel function of the first kind.

The Kaiser window function provides two improvements over the traditionally used Gabor stimulus [17]. First, the Kaiser window has higher values over a larger region than the Gaussian function used in the Gabor function. The greater stimulus width ensures that more of the test participant's field of view is filled with the periodic pattern, which is important as no headrest or viewing aids are used in the experiments. Second, the Kaiser window localizes more of the energy of the test signal in the main lobe, improving spectral specificity. In other words, the Kaiser window provides the largest stimulus with the smallest transition band [17] and, hence, optimised spatial frequency localization. The Kaiser window is multiplied by the sinusoidal pattern, providing a smooth transition from the pattern to the background region.

The Mira pixel brightness, normalized to the range [0,1], is defined by

$$M(x_0, y_0) = \begin{cases} \frac{1}{2}\left[\frac{I_0\left(\beta\sqrt{1-r^2}\right)}{I_0(\beta)}\sin\left(\frac{2\pi y_0}{T}\right)+1\right] & r\left(x_0, y_0\right) < 1 \\ \frac{1}{2} & \text{otherwise} \end{cases} \tag{3.5}$$

where T is the Mira period.

The highest measurable spatial resolution is determined by the period of the test signal, which cannot be less than twice the distance between two pixels of the display device in the direction where the specified period is minimal, according to the sampling theorem (45 degrees from the horizontal axis). In the experiments below, the display hardware limits the brightness values to 8-bit pixel values (0–black, 255–white) due to limited available equipment, but could be changed on more advanced equipment if designed. The pattern is presented in grayscale, and the background pixel brightness value is set to a nominal level. The following set of (8-bit) background pixel brightness levels are used in the experiments: $\{40, 80, 120, 160, 200\}$. Since the experiment uses the pixel brightness of the test stimulus and background brightness, the upper values of the equipment brightness of 255 are not acceptable in the experiment. For the entire stimulus to be assessed by the participants, the measurement interval is from 0 to 200.

Now let k be the spatial frequency, then the normalized spectrum $M(k_y)$ of the test object is given by,

$$M\left(k_y\right) = \frac{d\sinh\left(\frac{\pi}{\beta}\sqrt{\beta^2 - \pi^2 d^2\left(k-\frac{2\pi}{T}\right)^2}\right)}{2I_0\left(\beta\right)\sqrt{\beta^2 - \pi^2 d^2\left(k-\frac{2\pi}{T}\right)^2}}. \tag{3.6}$$

Naturally, the limits of spatial resolution are determined by the test object's width and the screen display's pixel pitch.

Temporal changes (flicker) in stimuli are introduced through a channel that is independent of the formation of the spatial component of the stimulus. The choice of liquid crystal technology enables the modulation of the backlight pixel brightness, hence control of temporal change [17]. An LG 22MP47A monitor (LG Corporation, Seoul, South Korea) is used. The pixel brightness

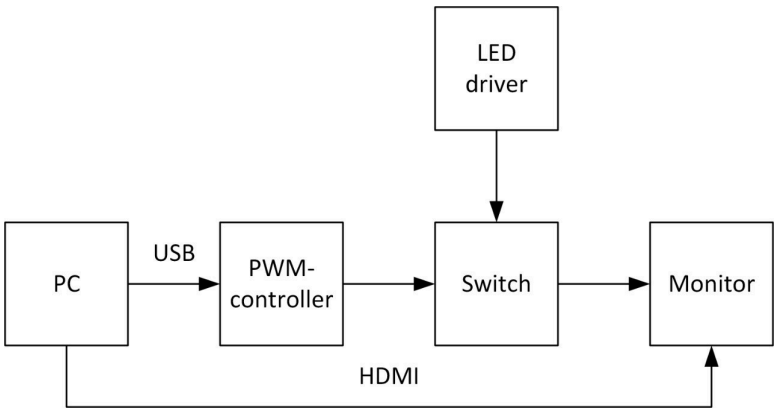


Figure 3.3: Structural diagram of the display system installation for research.

of the stimulus is sufficient to obtain representative results corresponding to daily viewing conditions. The switching nature of LEDs makes linear dimming of the monitor backlight LEDs impossible. Therefore, flicker via pulse-width-modulation (PWM) of the current flowing through the backlight LEDs (we also disabled the built-in PWM LED control of the monitor) is used. The PWM temporal modulation control blinks the entire display at the specified rate (see Fig. 3.3). PWM also provides near instantaneous response time of the flicker from the viewer's perspective (whereas each spatial pattern is static). The display's size and the participant's viewing distance allow us to assert that when the participant looks at the centre of Mira, the background region is not imaged in the zone of centre vision.

In experiments, the upper temporal frequency recorded by the eye was limited to 100 Hz, while in a limited range, it avoids discomfort and potential hazards such as epileptic seizures [120]. The PWM bandwidth of 3.9 kHz, much higher than any temporal change detectable by the HVS, ensures the flicker signal's accuracy. The PWM coefficient can represent the temporal component of the pixel brightness. Thus, visible pixel brightness is given by:

$$M(x, y, t) = m_{pwm}(t)M(x, y), \tag{3.7}$$

where $m_{pwm}(t)$ is the PWM signal of temporal frequency f Hz.

A photodetector was used to confirm the correctness of the temporal sig-

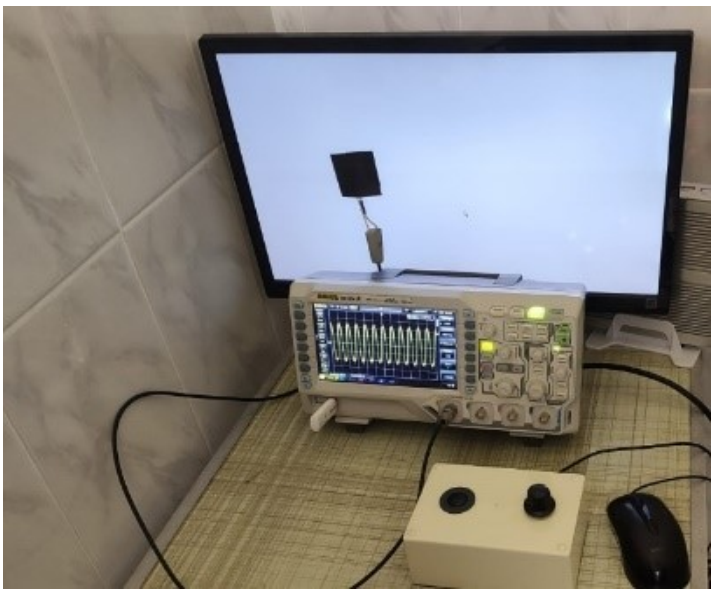


Figure 3.4: Photocell with a linear light-signal characteristic connected to an oscilloscope.

nal as it emerged from the screen, see Fig. 3.4. An X-rite i1 Pro (X-Rite Inc., USA) device was used to test the monitor color profile (d65). The monitor was calibrated according to the recommendation of the International Telecommunications Union [121]. These experiments are designed to quantify the perception of modern multimedia content, that is the standard gamma value for video transmission systems of 2.4 (Rec. ITU-R Bt.601-4 standard) is used. Displays are specified in terms of luminance, and consequently, knowing the form of the visibility model for a particular adapting luminance is needed when specifying rendering limits in space and time [87]. The results do not depend on the display size if the parameters during measurement are sufficient, where the pixel is smaller than the distinguishable object, and the stimulus covers the fields of clear vision. The pixel size determines the minimum distance, and the maximum distance is determined so that the stimulus does not become smaller than the clear-vision field. Hence, it is apparent that measurements on different types of screens are unnecessary; the result will be universal for 4K, 5K, and mobile technology. The diameter of the stimulus is 0.2 m. The maximum angular resolution of the human eye is one arc minute [84]. The pe-

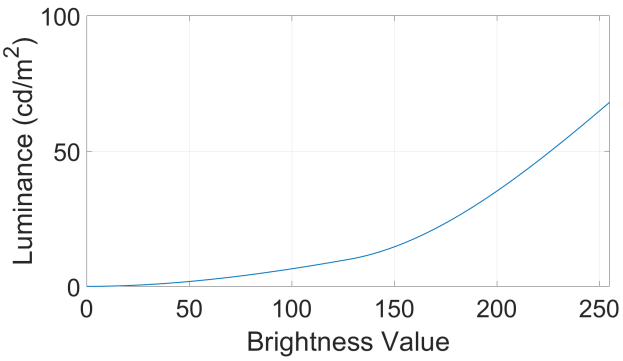


Figure 3.5: Dependence of luminance in the immediate vicinity of the monitor on the pixel brightness value.

riod at the highest spatial frequency has 2 pixels, and the lowest is 540 pixels. According to these data, the minimum allowable distance from the monitor for the stimulus to be in the centre vision zone is 0.872 m, and the maximum length is 1.149 m [115].

The dependence of luminance in the immediate vicinity of the monitor on the pixel brightness value (Fig. 3.5) is determined by adding to the experiments the relationship between backlight pixel brightness, retinal illumination (E_e), and pupil size (d_e). The possibility of understanding the interconnections of the effect of pupil size and ambient light level is also directly related to the problem of multimedia content and the problem of the clinical community [122]. The eye focuses light from one point in the world to one place (subject to the diffraction limit of the pupil) on the retina, thus forming an image on the retina that is a two-dimensional replica of the illuminated surfaces within the field of view [123].

The luminous flux through the participant's pupil was controlled during the experimentation. For this, the following parameters were determined using a lux meter: luminance at the point of eye position and near the monitor (Luminance), as well as the size of the pupil. The pupil of the human eye varies in size from a minimum of 2 mm at very high brightness to a maximum approaching 8–9 mm [124]. Psychophysical contrast sensitivity studies usually use a fixed artificial pupil [125] with a typical size of 2-3 mm [100, 126, 127].

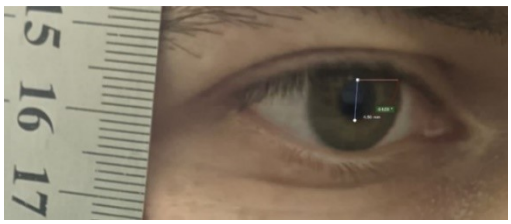


Figure 3.6: Pupil of the participant at pixel brightness 128.

However, in real life, both the pupil's size and the retina's illumination change along with the change in the environment's illumination level; the neural and optical components change depending on the size of the pupil [122]. To determine the size of the pupil, a photo of the participant's eye at the given monitor brightness, was taken after which the size of the pupil was analyzed, Fig. 3.6.

Referring again to Fig. 3.5 shows how it is possible to link the pixel brightness values used in the proposed work with the generally accepted brightness in cd/m (the candela, symbol cd, is the SI unit of luminous intensity in a given direction) and Troland (unit of conventional retinal illuminance), since $1\,\mathrm{cd/m^2} = 92$ Tr. Considering the pupil's adaptation, the luminous flux slightly decreases with increasing pixel brightness from 0.58 to 0.47 mm.

3.2 Procedure

The study was approved by the HECS Human Ethics Committee at the University of Waikato (HECS-20-64, HECS-20-58).

Forty-two participants aged 20 to 40, with uncorrected vision (glasses, lenses), were recruited through the Moscow Technical University of Communications and Informatics. In this experiment, normal vision is determined by typical user-generated content. The participants do not use glasses, lenses, or other medical devices to correct their vision in normal daily activities, and the participants are free of known neurological disorders. Thirty-eight of the participants (90%) have no experience working with human perception of visual information. Informed consent was obtained from all participants. Most

participants were third-year undergraduate students, which is a good balance between three important parameters: physical maturity of the eye, daily use of typical user-generated content, namely video and image viewing on the internet, and lack of experience with work visual information perception. Of these three parameters, lack of experience with visual perception is especially important, as having such experience leads to improvement in artefact detection and spatial-temporal threshold detection [17].

The participant begins the test by pressing one of the buttons of a three-button mouse. The participant was provided a remote control with a rotatable controller and a button. The rotating control made the stimulus brighter. The participant rotated the manipulator control to the minimum level at which the stimulus could be discerned, and after that, the participant pressed the button. Upon pressing the button, the stimulus level value was recorded.The participant is not limited in time to view the test image, ensuring chromatic adaptation (the ability of the human visual system to adapt to changes in lighting to preserve the appearance of the colours of objects). As the tests were carried out at an individual pace, the next pattern was not presented until the participant recorded the answer by pressing a button on the remote control. The tests were presented to the participants in random order. The experimented situation is intended to simulate a normal environment for content consumption closely. Hence, these experiments represent a non-classical approach. Consequently, no head restraints or viewing aids such as lenses are used. Computer-assisted testing was conducted in a room without natural light using two light sources, one of which was the display, which is consistent with typical user content. The same light was provided for all participants. Participants were instructed to maintain a certain distance from the monitor.

The trials were undertaken at the participant's own pace in an effort to reduce the effects of fatigue, and the participant was allowed to take a rest break at any time. A 95% confidence interval for our data was used. The standard deviation to evaluate the confidence interval for each presentation

is given in the Rec. ITU-T Bt.500-15 standard [28]. All experiments were continued until the confidence interval fell below 5% of the current mean value for each point.

3.3 Temporal Contrast Sensitivity

3.3.1 Subjective Test

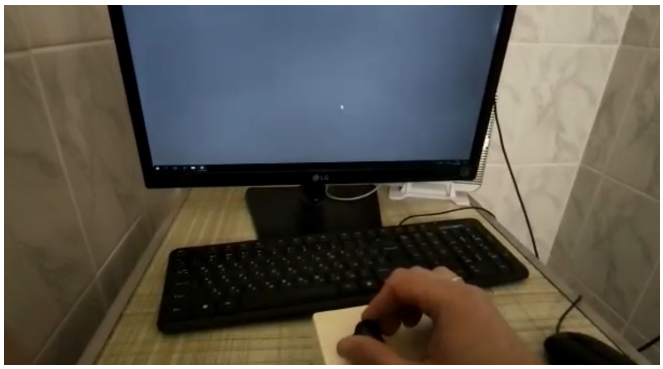


Figure 3.7: Subjective test with temporal flicker.

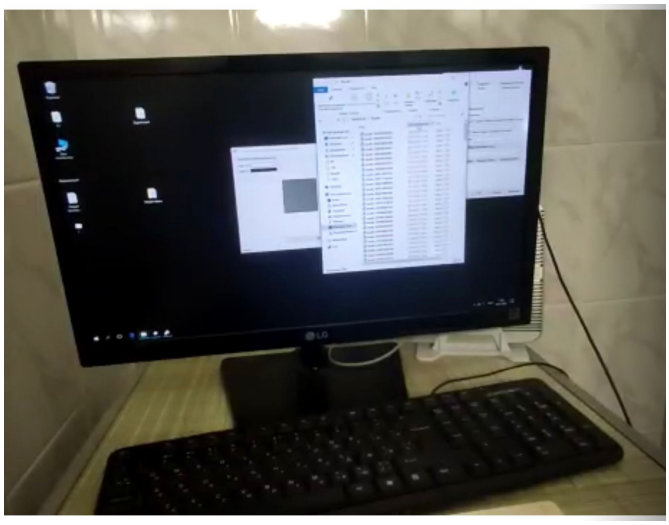


Figure 3.8: Software recorded thresholds.

A uniform grey screen (i.e. background only) with temporal flicker is presented to the participant, who then regulates the flicker amplitude to the minimum noticeable level, which is recorded in Fig. 3.7, Fig. 3.8. Contrast

Table 3.1: The collection number of evaluations for the experiments.

Experiment	Temporal Contrast Sensitivity	Spatio-Temporal Contrast Sensitivity
Number of measured temporal frequencies	12	12
Number of measured spatial frequencies	-	8
Number of pixel brightness stimulus levels	5	5
Number of evaluations of one participant	60	960 (double evaluation of each stimulus, first spatial, then temporal)
Number of participants	13	29
Total number of evaluations of experiment	780	27840

thresholds were measured at 12 different temporal frequencies and five different background levels. The following set of (8-bit) background pixel brightness levels are used in the experiments: $\{40, 80, 120, 160, 200\}$. In this experiment, 13 participants (5 women and 8 men) each recorded 60 levels. In total, 780 measurements were obtained (from all of the first experiment participants), Table 3.1. The test took about 15 minutes per participant.

3.3.2 Results

The results of the experiment's temporal (flicker) frequency are presented in Fig. 3.10 and Fig. 3.9. The linear fit of the contrast threshold with temporal frequency is given, where and throughout the text, the log(s) is the decimal logarithm.

When compared with Watson's work, namely the graphs with de Lange's results [87], the model of temporal contrast sensitivity has a similar shape as

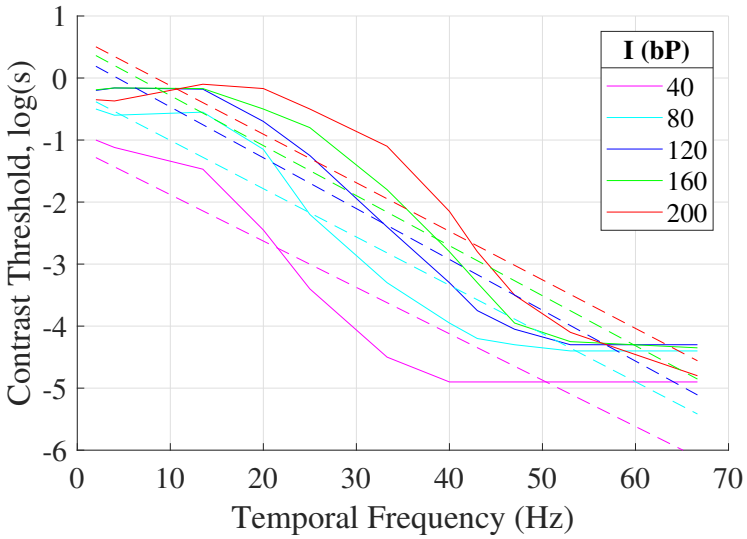


Figure 3.9: The results for the temporal components of responses of the participants in the experiment compared to the linear model, illuminance, is presented in brightness background pixel (bP). Dashed lines are the linear model. Solid lines are from measured data.

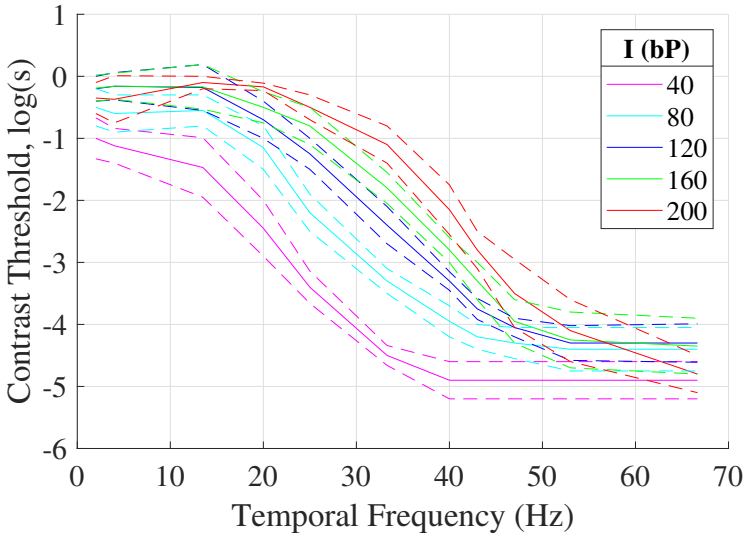


Figure 3.10: The results for the temporal components of responses of the participants in the experiment, illuminance, is presented in brightness background pixel (bP). Dashed lines are the confidence interval. Solid lines are from measured data.

previous research in which old technology (CRT) was used, Fig. 3.11. Watson gets a slope of -0.064 for the log contrast sensitivity against temporal fre-

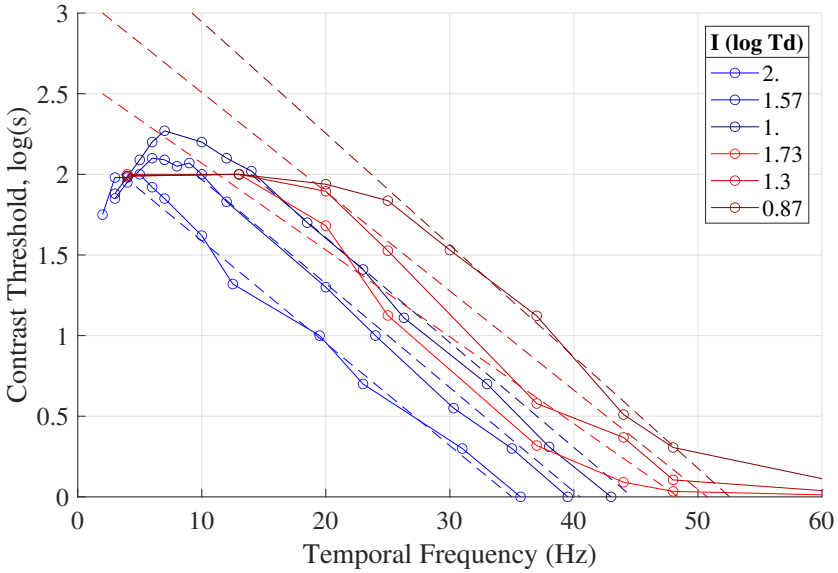


Figure 3.11: The results for the temporal components of responses of the participants compared to the linear model and Pyramid of Visibility. Dashed lines are the linear model, and solid lines are from the measured data. Following Watson's work, illuminance is presented in Troland [87]. Red: this study. Blue: Watson's work.

quency. In this study, the slope is -0.058. However, the fit is only good for a narrow range, where the range of good fit depends on illuminance.

3.4 Spatio-Temporal Contrast Sensitivity

3.4.1 Subjective Test

At first, the Mira pattern is not visible. The participants controlled the flicker amplitude, setting it to the minimum level at which they could discern the Mira pattern. Then, participants adjust the amplitude of the Mira pattern until they can minimally distinguish it from the background. Contrast thresholds were measured at eight different spatial and twelve temporal frequencies at five different background pixel brightness levels. Each of the 29 participants (5 women and 24 men) performed 960 threshold evaluations. A total of 27840 estimates were acquired from the second experiment, Table 3.1. The testing

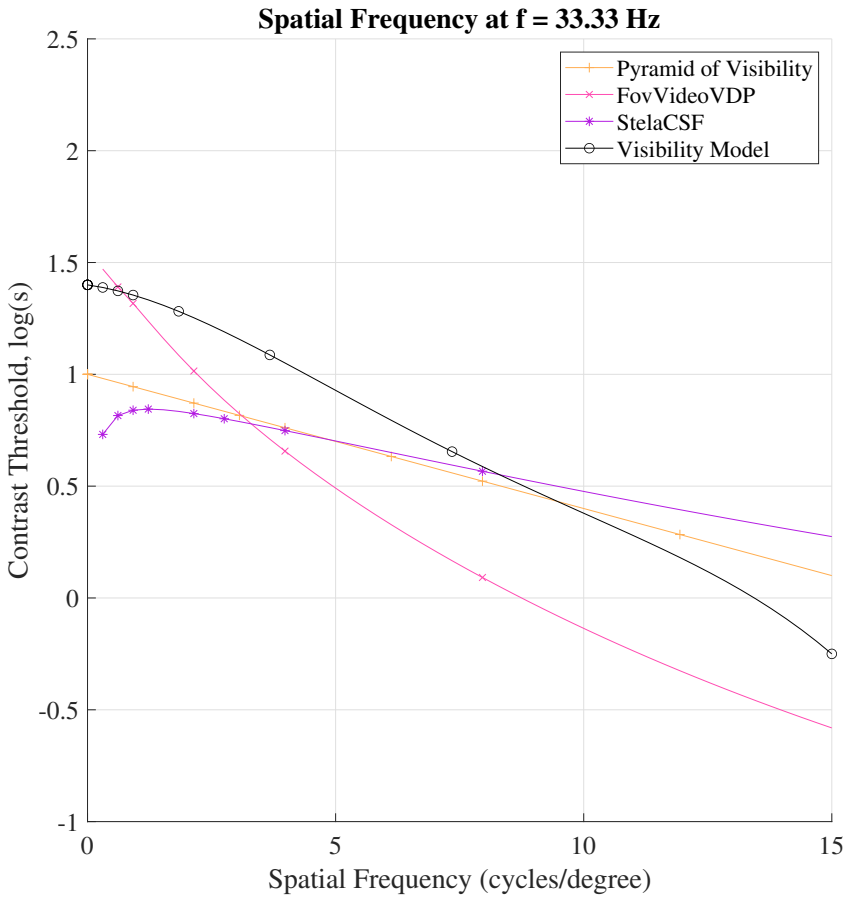


Figure 3.12: Contrast sensitivity as a function of spatial frequency at a temporal frequency from experiment compared to the Pyramid of visibility [87], FovVideoVDP [4], stelaCSF [91], f=33.33 Hz. Yellow: Pyramid of visibility. Purple: FovVideoVDP. Red: stelaCSF. Black: this study.

for the second experiment was about 30 minutes per participant.

3.4.2 Results

Kulikowski and Watson's work, as shown earlier, over a broad and important range of luminance, spatial frequency, and temporal frequency, the sensitivity can be described by a linear model [99, 87]. In contrast, in this Chapter, representing the temporal characteristics of the visual system using linear functions under the conditions of the experiment is only possible within a limited interval of values of the temporal frequency, Fig. 3.12, Fig. 3.13, Fig. 3.14, Fig. 3.15. This nonlinearity is in contrast to the earlier work of Kulikowski,

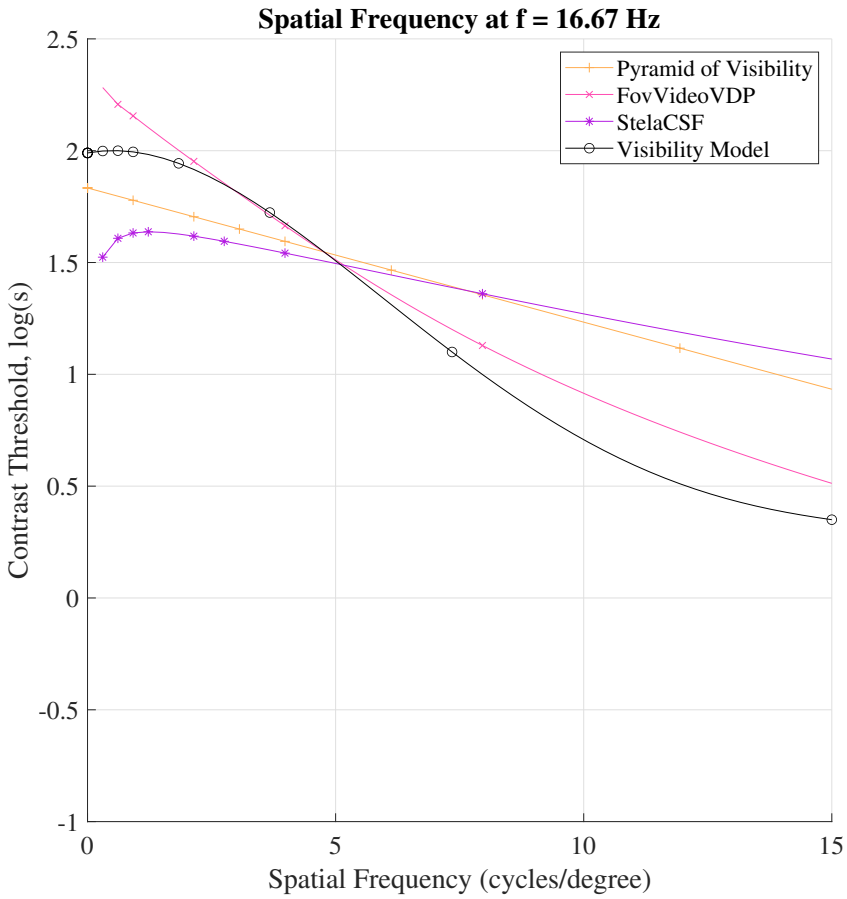


Figure 3.13: Contrast sensitivity as a function of spatial frequency at a temporal frequency from experiment compared to the Pyramid of visibility [87], FovVideoVDP [4], stelaCSF [91], f=16.67 HZ. Yellow: Pyramid of visibility. Purple: FovVideoVDP. Red: stelaCSF. Black: this study.

who intervened with feedback to stabilize the observed visual system parameters [99], and Robson, who used binoculars [107]. This study employs neither measure. An explanation for this difference and peak at 8 Hz needs to be more forthcoming and is a topic for future research. It can be assumed that such a difference in comparison with previous works is because the earlier experiments had independent goals of separately detecting the spatial and temporal threshold. Moreover, each spatial and temporal component was fixed in Watson's work, and the other varied [87]. The work presented herein employed viewing conditions more similar to real-world media consumption (hardware, no head holder).

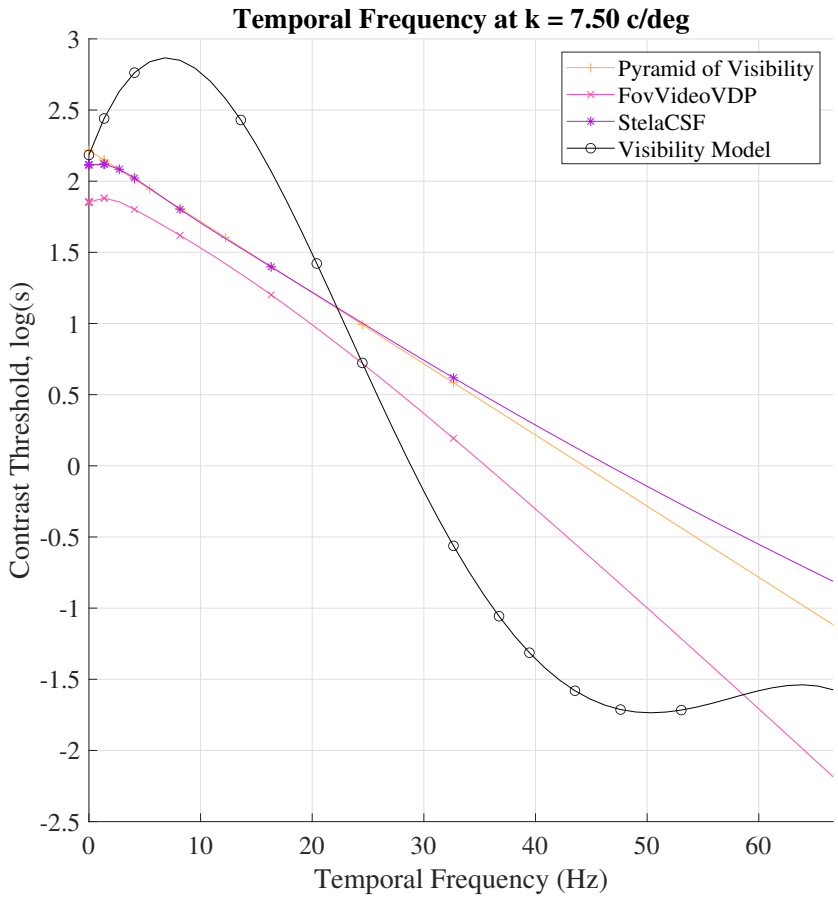


Figure 3.14: Contrast sensitivity as a function of temporal frequency at spatial frequency from experiment compared to the Pyramid of visibility [87], FovVideoVDP [4], stelaCSF [91], k=7.5 c/deg. Yellow: Pyramid of visibility. Purple: FovVideoVDP. Red: stelaCSF. Black: this study.

3.5 Contrast sensitivity and background pixel brightness

Figs. 3.16 and 3.17 show graphs of the spatio-temporal contrast sensitivity of the subject's visual perception for four pixel-brightness background levels. An explicit dependence of the temporal and spatial contrast sensitivity on the pixel brightness at high spatial frequencies is shown.

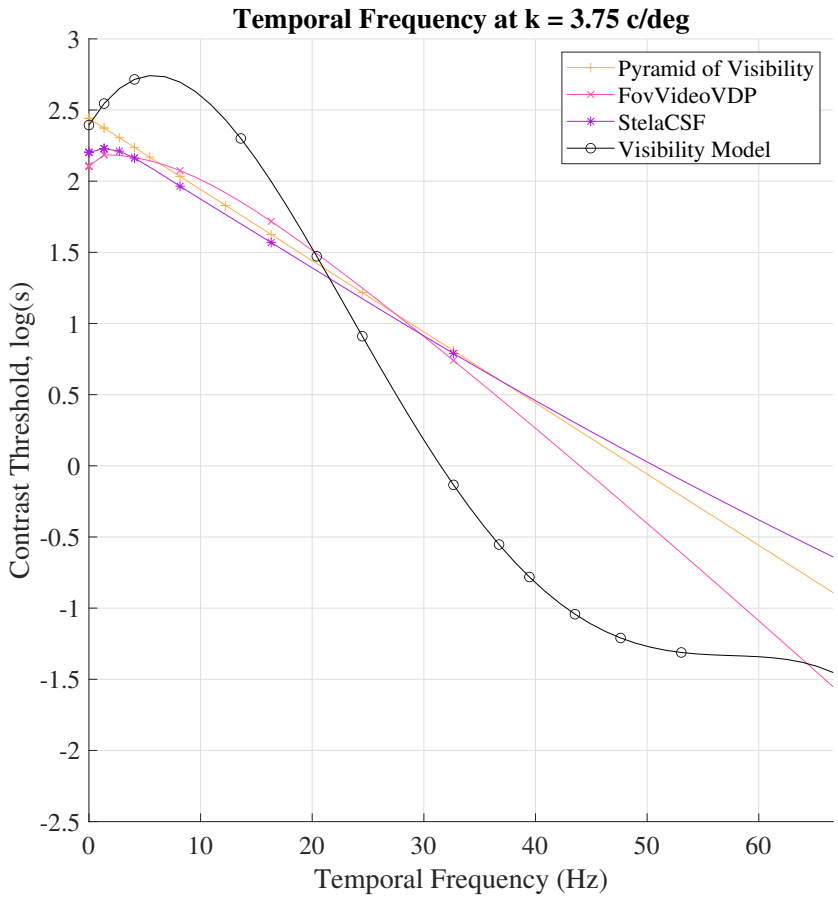


Figure 3.15: Contrast sensitivity as a function of temporal frequency at spatial frequency from experiment compared to the Pyramid of visibility [87], FovVideoVDP [4], stelaCSF [91], k=3.75 c/deg. Yellow: Pyramid of visibility. Purple: FovVideoVDP. Red: stelaCSF. Black: this study.

3.6 Analysis

The results presented herein describe the spatial and temporal frequency response of the HVS over a broad range of relevant values. While recent evidence exists that the HVS can perceive temporal changes up to 500 Hz under controlled conditions, the results, under conditions more closely resembling real-world content consumption, showed that HVS no response of more than 200 Hz flicker [129]. The bandwidth of the equipment is 3.9 kHz, which is faster than humans perceive flicker artefacts [129]. Therefore, the principal source of measurement error is (as elsewhere) imperfection in the test apparatus and the measurement process. It should be noted here that, as shown above, the

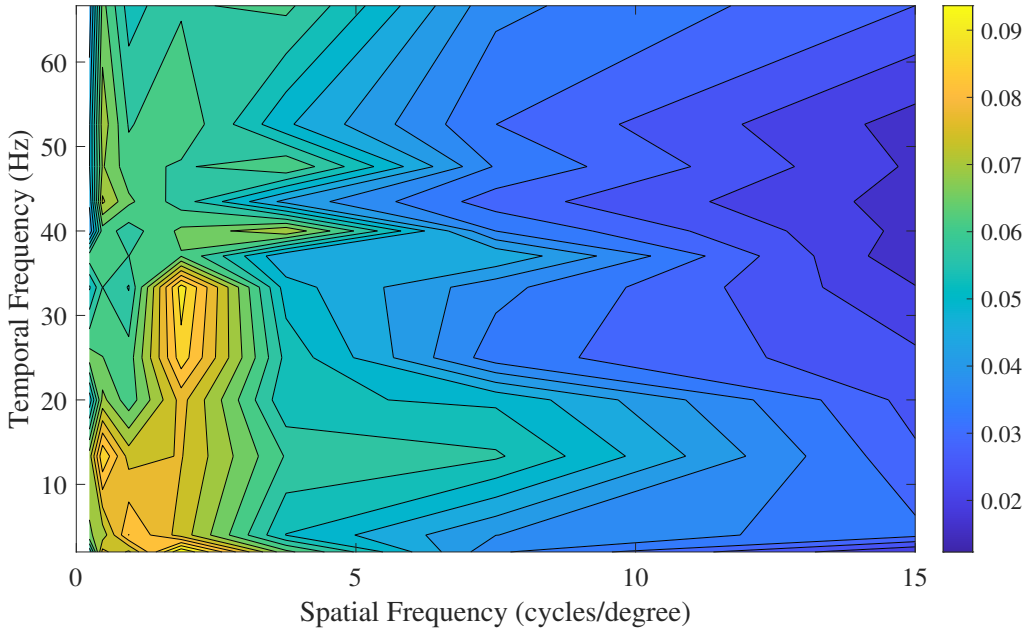


Figure 3.16: Graph of the spatio-temporal contrast sensitivity of the participant's visual perception for pixel brightness 200.

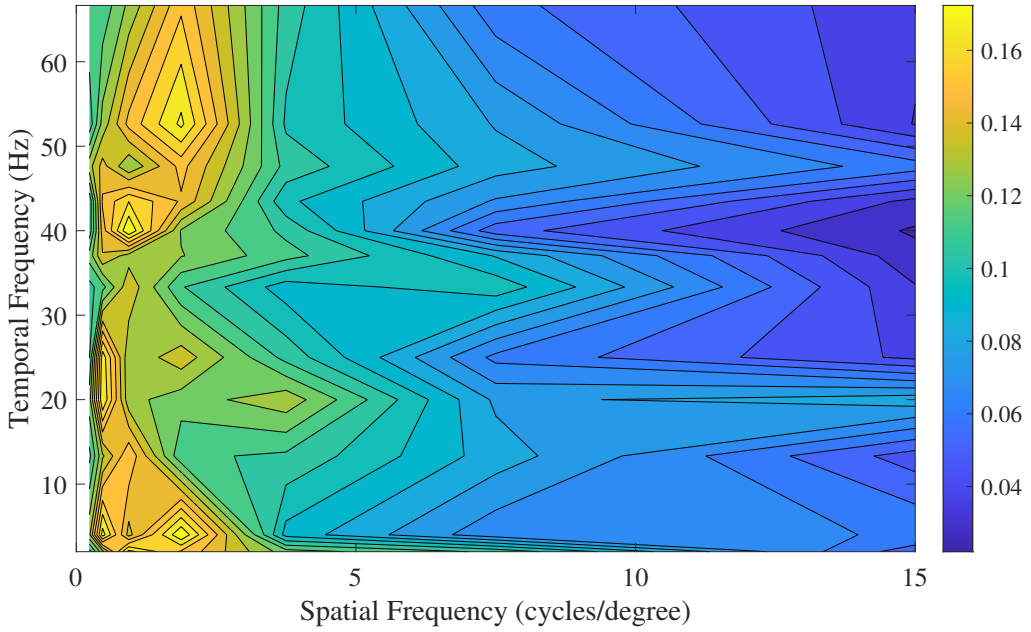


Figure 3.17: Graph of the spatio-temporal contrast sensitivity of the participant's visual perception for pixel brightness 80.

equipment previously used to perform measurements, particularly the generation of stimuli, was very imperfect. At the same time, the imperfection of the current equipment must also be noted. Despite the use of an IPS matrix, which is characterized by significant inertia, the control features of its elements,

in combination with modulation of the backlight brightness using PWM, can create unpredictable spurious maxima of the spatio-temporal spectrum of the displayed stimulus. Analysis of the results shows that participants noticed flickering at different relatively high frequencies at different pixel brightness values, which is reflected in the graphs as bursts of temporal sensitivity, for example, at 15 and 40 Hz (Fig. 3.16), at 20 and 48 Hz (Fig. 3.17) respectively, at 200, 80 levels of pixel brightness.

The control features and elements of an IPS matrix monitor, in combination with PWM modulation of the backlight pixel brightness, can create spurious maxima of the spatio-temporal spectrum of the displayed stimulus. The results show that participants noticed flickering at the higher frequencies tested. In other words, it saw these spurious maxima, causing unexpected changes in the threshold values. These are excluded from processing the results (nevertheless, these exclusions caused very little change in the results). Prolonged display flickering at high amplitude may trigger seizures in people not diagnosed with photosensitive epilepsy. Hence, reverse threshold measurements for participants transitioning from high modulation to zero were not performed (clinical practice shows that 76 percent of photosensitive people do not know about their photosensitivity [130]). Also, as the goal was to bring the experiments as close as possible to the typical content consumption, a head holder was not used. Regardless, participants were asked to keep their heads within a fairly narrow range of positions. Consequently, any head movement is compensated for by large-scale experiments and significance testing: 960 thresholds are found in the experiment for one participant, terminating when significance is obtained.

3.7 Summary

This chapter presented a novel concept for measuring contrast sensitivity function. The proposed concept was developed into a working proof-of-concept

equipment. Additionally, in this chapter, the 27840 thresholds of the visibility of spatio-temporal sinusoidal variations necessary to determine a human's perception of artefacts were measured by a new method using different spatial sizes and temporal modulation rates. The presented measuring data, in turn, provides a new refined visibility model as there were no experiments previously without medical intervention and considering typical user content.

Chapter 4

Modelling the Human Contrast Sensitivity Function

Watson determined the vision model as follows:“ A model is defined as a simulation of some physical system. Vision science has numerous modest, ad hoc models of small performance components, mostly in a form less explicit than computer code” [83]. A contrast sensitivity function (CSF) is a filter that determines the neuron response in the early visual system due to visual stimulus. The CSF model is the first part of early vision, namely that of the filtering stage, which governs what spatiotemporal fluctuations in stimuli the HVS responds to.

This chapter presents the refinement of the multidimensional model of human contrast sensitivity parameters of video transmission when viewed with modern display technology. Since there are very few video evaluation metrics based on fundamental knowledge of the operation of the human visual system, a new full-reference metric is created to test the model, expanding the metric of the peak signal-to-noise ratio by the characteristics of the CSF. The results of testing the proposed model and comparison with existing visibility models are presented and analyzed. The results are obtained by including the developed CSF model and existing CSF models into two video quality metrics. This video quality metrics test on three independent public datasets.

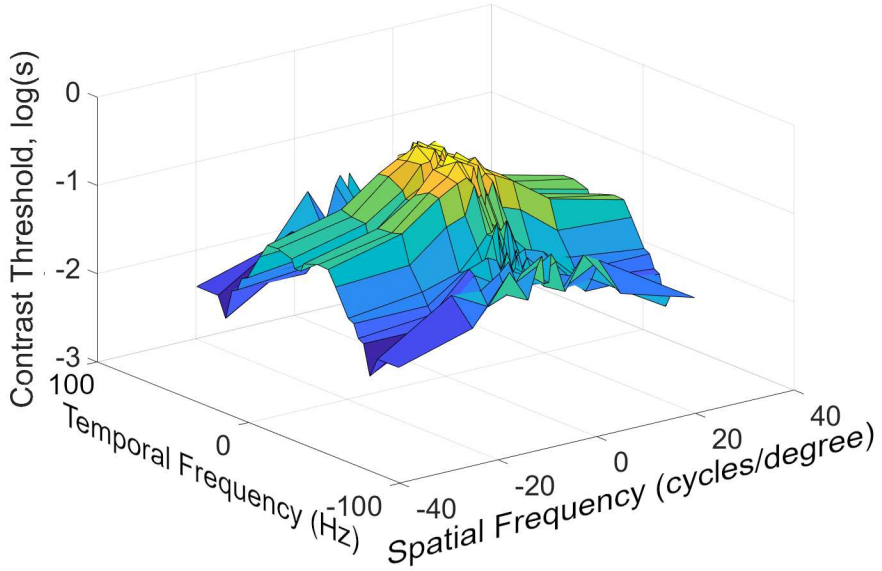


Figure 4.1: The average of measured values for all evaluations of the contrast threshold of the experiment on 120−pixel brightness level.

The work in this Chapter was presented in [9, 60, 61, 134] during this doctoral research project.

4.1 The Model of Visibility

Fig. 4.1 shows average real values for all scores of the spatial-temporeal measuring experiment in Chapter 3 at a pixel brightness stimulus level of 120.

An important result of Chapter 3 showed that a linear Watson's pyramid of visibility model may be appropriate for spatial frequency dependence but is not for temporal frequency. Therefore, a polynomial approximation of the model based on the results from the subjective experiment is hereby developed. The model is found by linear regression against polynomial terms. By inspection, it is found that a 4th-order polynomial strikes an appropriate balance between encompassing non-linearity and minimising over-fitting. The model is,

$$\log(s)\,(k', f', l') = \sum_{\alpha,\gamma,\delta:\alpha+\gamma+\delta\leq 4} c_{\alpha,\gamma,\delta} k'^{\alpha} f'^{\gamma} l'^{\delta}. \tag{4.1}$$

where

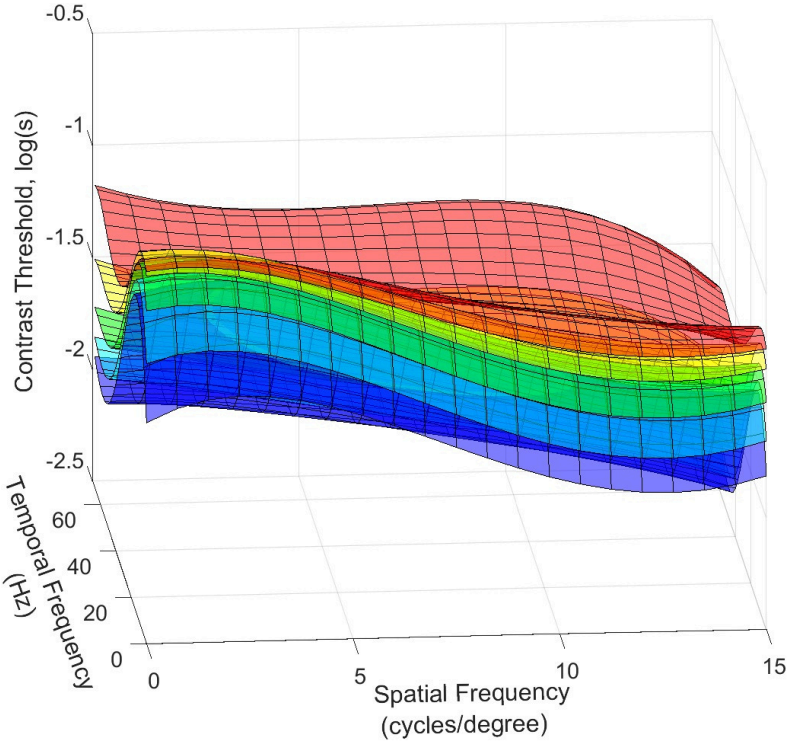


Figure 4.2: The model of visibility is built from the values of Experiment 2 (Section 3.4, Page 59)

with a background pixel brightness stimulus level of red − 40, yellow − 80, green − 120, blue − 160, purple − 200, where s=0.1.

$$k^{'} = \frac{k}{15} \quad , \quad f^{'} = \frac{f}{66.6} \quad , \quad l^{'} = \frac{l}{200} \quad , \tag{4.2}$$

are normalized parameters setting such that $k^{'}, f^{'}, l^{'} \in [0, 1]$, that is, the maximum spatial frequency in the experiment is 15 cycles per deg, the maximum temporal frequency is 66.6 Hz, and the maximum pixel brightness is 200. Eq. 4.1 may be written in the following matrix form

$$\log(s)(k^{'}, f^{'}, l^{'}) = \mathrm{Xc}^{\mathrm{T}}, \tag{4.3}$$

where, now, $\log(s)$ is a vector of the logarithm of the measured s values.

The $c_{\alpha,\gamma,\delta}$ in Table 4.1, viz.:

$$c = \left(\mathrm{X_i^T X_i} + \lambda I^{'}\right)^{-1} \mathrm{X_i^T} s, \tag{4.4}$$

where λ is a regularization coefficient with $\lambda = 0.001$, $I^{'}$ is a modified identity

matrix such that $I^{'}(0,0) = 0$, , and each row of the X_i matrix is the X matrix with combinations of l, f, k for the i^{th} experiment.

Table 4.1: The coefficients of the approximation polynomials are calculated for $l = 120$.

I	c_i	I	c_i	I	c_i	I	c_i
1	2.3132	$k^{'3}$	8.1636	$k^{'4}$	2.5845	$k^{'2}l^{'2}$	3.5279
$k^{'}$	1.5051	$f^{'3}$	10.6719	$f^{'4}$	7.1957	$f^{'2}l^{'2}$	1.9792
$f^{'}$	0.3673	$l^{'3}$	3.4370	$l^{'4}$	7.1957	$k^{'2}f^{'}l^{'}$	2.0704
$l^{'}$	3.1866	$k^{'2}f^{'}$	26.5144	$k^{'3}f^{'}$	9.5953	$k^{'}f^{'2}l^{'}$	5.3859
$k^{'2}$	7.6661	$k^{'2}l^{'}$	2.3375	$k^{'3}l^{'}$	3.8210	$k^{'}f^{'}l^{'2}$	0.6535
$f^{'2}$	0.6479	$k^{'}f^{'2}$	5.9001	$k^{'}f^{'3}$	0.1436	-	-
$l^{'2}$	11.7062	$f^{'2}l^{'}$	29.0300	$f^{'3}l^{'}$	17.8684	-	-
$k^{'}f^{'}$	3.8606	$k^{'}l^{'2}$	6.5085	$k^{'}l^{'3}$	0.7139	-	-
$k^{'}l^{'}$	1.4359	$f^{'}l^{'2}$	14.2167	$f^{'}l^{'3}$	4.2082	-	-
$f^{'}l^{'}$	6.8897	$k^{'}f^{'}l^{'}$	2.5296	$k^{'2}f^{'2}$	1.3122	-	-

In Fig. 4.2, the system model for different background pixel brightness stimulus levels is shown, demonstrating that, while the overall contrast threshold changes with pixel brightness, the shape of the contrast threshold characteristic depends little on pixel brightness. As the temporal frequency increases, the spatial frequencies that are easier to see become less noticeable more quickly than those that are more difficult to distinguish from the background. At less noticeable spatial frequencies, the temporal frequency is the dominant factor in terms of decline in the contrast threshold, in agreement with much earlier work [120].

4.2 New Full Reference Video Metric Considering the Features of the Human Visual System.

Nowadays, numerous video quality metrics are available. However, only a few video metrics model the human visual system, namely the adaptation of the human visual system, spatial, temporal and peripheral aspects. Here, a new quality metric is proposed that extends the peak signal-to-noise ratio metric with features of the human visual system measured as described in Chapter 3. This and other commonly used quality metrics are compared by Pearson's linear correlation coefficient of the various video compression quality metrics with human subjective scores on videos from the publicly available Netflix and MCL data sets.

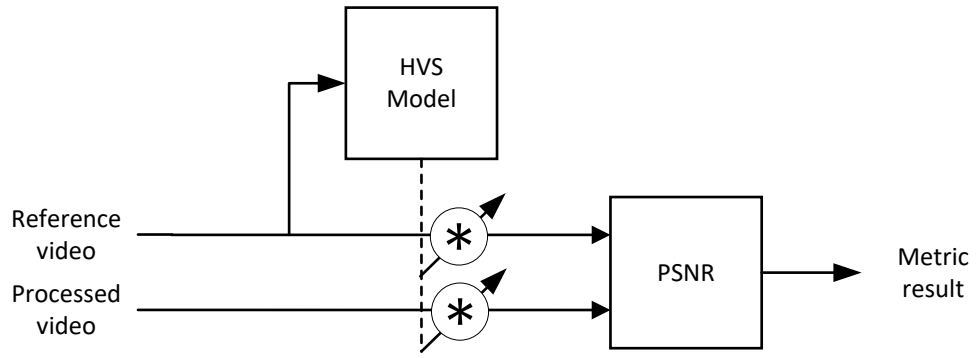


Figure 4.3: Block diagram PSNR-M+ inclusion of HVS model, where $*$ is the convolution, the arrow behind convolution indicates the change in the signal level according to the HVS characteristics.

4.2.1 Spatio-Temporal Component

The proposed full-reference method (PSNR-M+), see Fig. 4.3, is centred around a weighted PSNR calculation given by,

$$\mathrm{PSNR}'(I(t), I_R(t), t) = \mathrm{PSNR}(I(t)K(t), I_R(t)K(t), t), \tag{4.5}$$

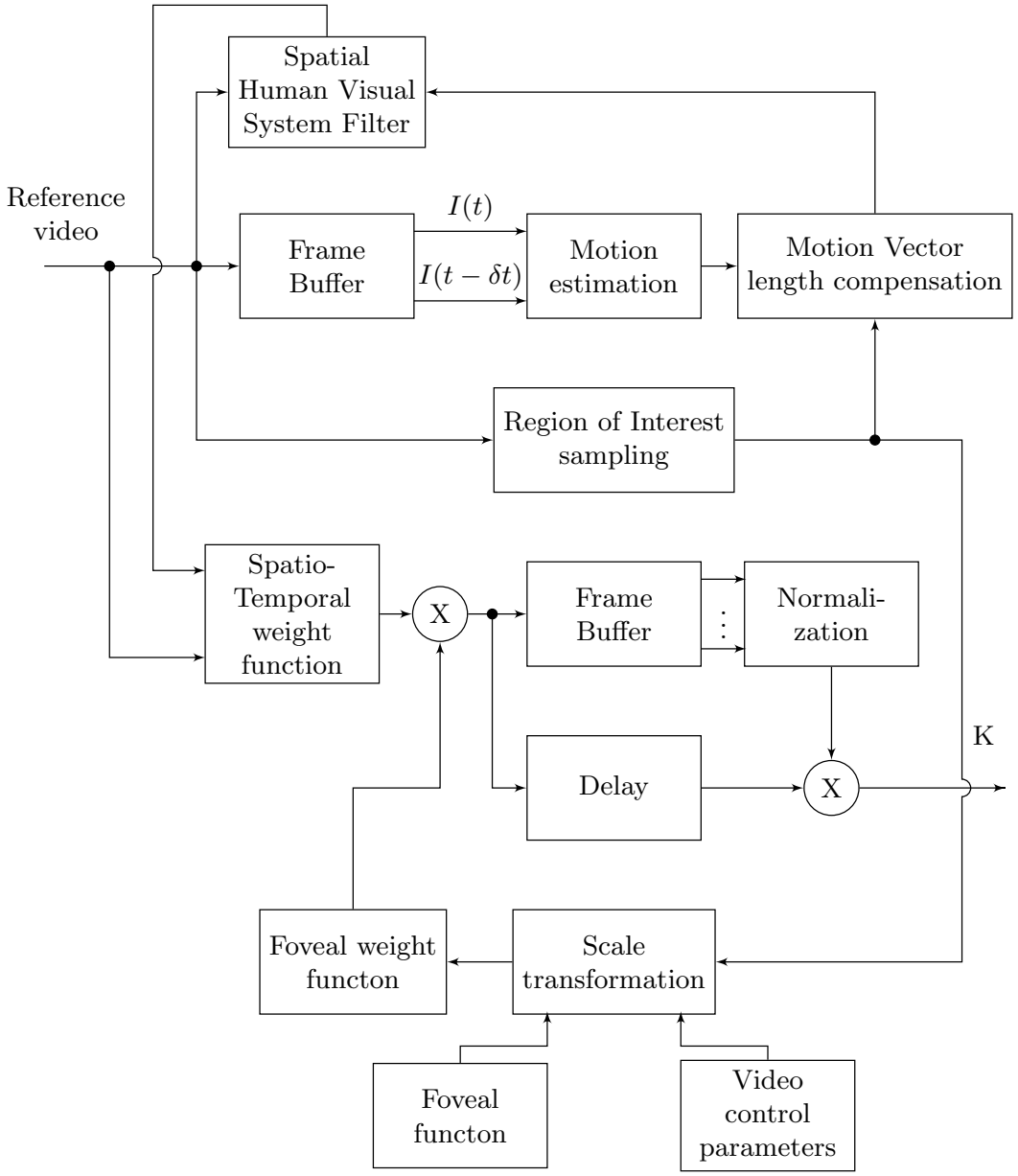


Figure 4.4: The framework of the methodology for weight estimate.

where I is a compressed frame, I_R is the reference uncompressed frame, and $K(t)$ is a weight coefficients matrix for t frames. The quality of the distorted video is measured by incorporating both the spatio-temporal-luminance component and a peripheral component. A flow diagram of the methodology for calculating $K(x, y, t)$ is shown in Figs. 4.4 and 4.5.

In the spatio-temporal block of Fig. 4.4, we use the weight function $H_{L,f_t}(f_x, f_y)$, where f_x, and f_y are the spatial frequencies, f_t is the temporal frequency, and L is the luminance [9]. This weight function models the ability

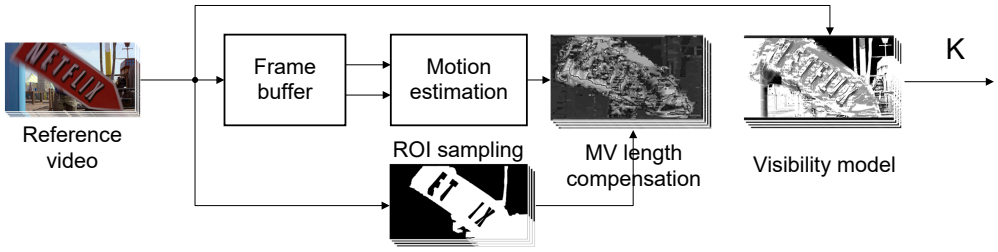


Figure 4.5: Flow diagram framework of the methodology for calculating $K(x, y, t)$.

of the human visual system to respond to spatio-temporal change, as via the CSF measured in the spatiao-temporal experiment of Chapter 3 (Section 3.4, Page 59). Note that,

$$h_{L,f_t}(x, y) \overset{F}{\Longleftrightarrow} (H_{L,f_t}(f_x, f_y)), \tag{4.6}$$

where $\overset{F}{\Longleftrightarrow}$ is the (invertible) Fourier transform. Assuming the HVS is isotropic in spatial frequency [131], then we may reduce the above to

$$f_0 = \sqrt{f_x^2 + f_y^2}. \tag{4.7}$$

Then,

$$H_{L,f_t}(f_x, f_y) = M(L, f_t, f_0), \tag{4.8}$$

where M is the spatio-temporal characteristic of the HVS. Filtering is performed in the spatial domain via.

$$I_R^{'}(x, y) = I_R(x, y) * h_{L,f_t}(x, y), \tag{4.9}$$

where $*$ is convolution. The spatio-temporal-luminance weighting factor, K_{stL}, is then computed as

$$K_{\mathrm{st}L}(x, y) = \frac{I_R^{'}(x, y)}{I_R(x, y)}. \tag{4.10}$$

The region of interest (ROI) sampling aims to identify objects that are more significant for the HVS. In this work, the ROI distinguishes individual objects using a variant of the watershed algorithm [132]. The watershed algorithm was chosen because of its flexibility and adaptability to various data types.

No more than five objects are selected close to the centre, with fewer objects selected if the total area of the selected clustered object is greater than a user-determined threshold [9].

The motion estimation utilises an adaptation of the MPEG block matching technique [131], which uses a 16×16 pixel block size method with $32\times$ 32-pixel search area. The motion vectors, $\underline{v}$(x,y,t), are compensated by subtracting the average within the ROI, v_{ROI}, that is,

$$v_{\text{comp.}}(x, y, t) = v(x, y, t) - v_{\text{ROI}}, \tag{4.11}$$

4.2.2 Peripheral component

The peripheral coefficient is an approximate foveation function, which models the decrease of focus from the centre of the ROI outwards. Foveation is the blurring which increases from the centre of vision, which, in the metric, is accounted for by assigning different weights to the central and peripheral areas. Since the user can look at different places on the screen depending on the ROI, we first find the ROI's centre to find the viewing angle.

$$x_c = \frac{\sum_{i=1}^{n} x_i}{n}, \quad y_c = \frac{\sum_{i=1}^{n} y_i}{n}, \tag{4.12}$$

where x_i and y_i are the coordinates of the i-th point of the region of interest.

Then, assuming that the user is at such a distance that there is 1 pixel in the centre of the screen that corresponds to 1 minute of arc of vision [5], the viewing angle for the pixel as

$$\alpha(x, y) = \arctan\left(\frac{\pi\sqrt{(x - x_c)^2 + (y - y_c)^2}}{1080}\right); \tag{4.13}$$

where x_c, and y_c are the centre of the ROI, and 1080 is the number of pixels width of the screen.

For the foveation function, the characteristics of the spatial distribution of receptors are taken as described by Gonzalez and Woods [132], Fig. 4.6 and fitted an approximate function to that distribution:

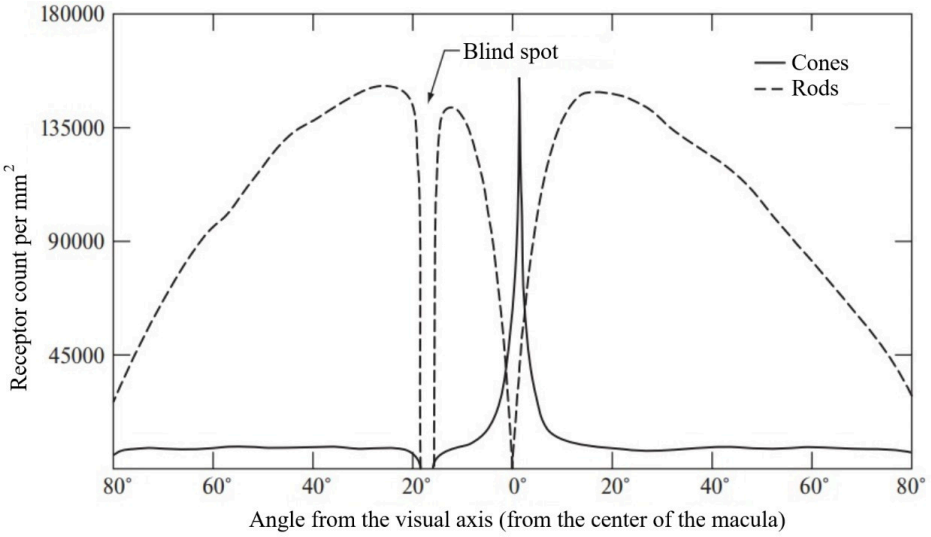


Figure 4.6: Distribution of rods and cones in the retina [132].

$$F(\alpha) = k_0 + k_1\alpha + k_2\alpha^2 + k_3\alpha^{k_4\alpha}, \tag{4.14}$$

where k_0 to k_4 are the coefficients at which this function is most similar to the original one [132]. The coefficients can be found by minimizing the standard deviation of the values of the approximating function from the original one, is, minimizing,

$$\Delta A = \frac{1}{N}\sqrt{\sum_i \Big(A(\alpha_i) - F_A(\alpha_i))^2\Big)} \tag{4.15}$$

where N is the number of values.

The approximation function (Gonzalez and Woods [132]) corresponds to :

$$A(\alpha) = \left(1.26 \times 10^5\right) e^{-0.71\alpha} + \left(1.5 \times 10^4\right) - 512\alpha + 11\alpha^2 - 0.08\alpha^3, \tag{4.16}$$

Note that the large numbers are normalised by the PSNR′ calculation below. The visual resolution decreases from the region of interest to the periphery according to the above expression [133]. Then, the peripheral coefficient in the weight function in the metrics is determined by

$$K_{pr}\,(x, y) = A(\alpha(x, y)). \tag{4.17}$$

In Fig. 4.7 we show an example of the $K_{\mathrm{pr}}(x, y)$ pattern.

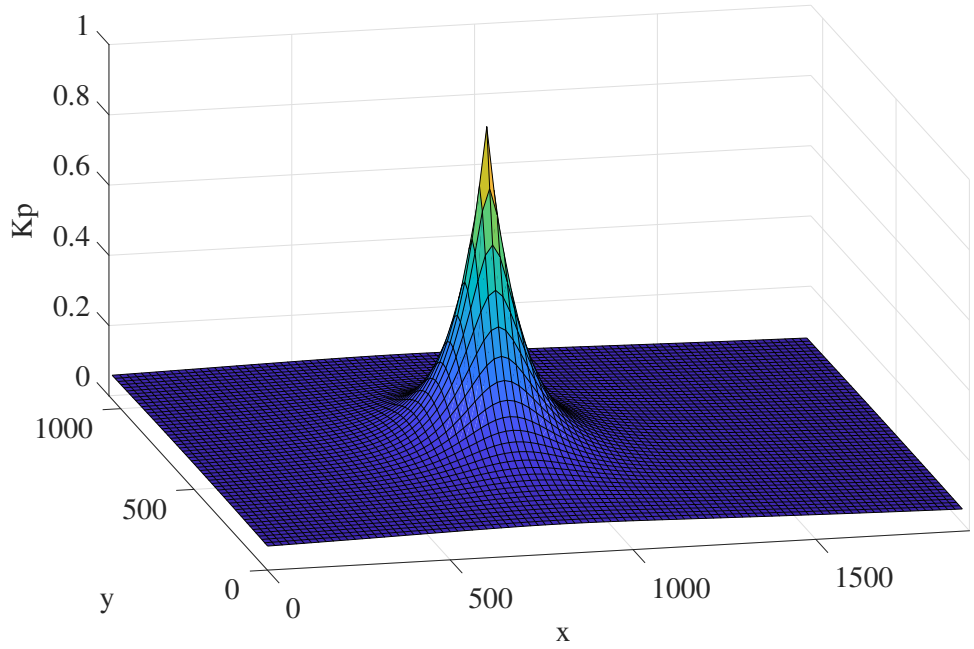


Figure 4.7: An example of K_{pr}, where the frame size is 1920 $\times$ 1080, and the centre of the ROI is set to (760, 640).

The methodology for weight estimation developed and presented above is introduced into the PSNR metric by weighting the function and the original video sequences via

$$\text{PSNR}' = 20\log_{10}\left(\frac{\sqrt{\sum_{t_n=-\frac{n}{2}}^{\frac{n}{2}}\sum_{x,y}K_{stL}^2\left(x,y,t_0+t_n\right)K_{pr}^2\left(x,y,t_0+t_n\right)}}{\sqrt{(n+1)\sum_{x,y}(I-I_R)^2K_{stL}^2\left(x,y,t_0\right)K_{pr}^2\left(x,y,t_0\right)}}\right), \tag{4.18}$$

where t_0 is the current frame, and t_n is the number of frames from the current frame, $n = \Delta t f_v$, Δt is the reaction time of the eye to frames (without considering the cognitive factors) [9], and f_v is the number of frames per second, $\Delta t = 0.8$ seconds [120]. According to the joint work of [134], there is a need for video quality metrics to contain a cognitive component (human perception of video) considering the adaptation time of the visual system to correctly model performance. Fig. 4.8 shows a block diagram of the PSNR-M+ metric with an enabled adaptation module.

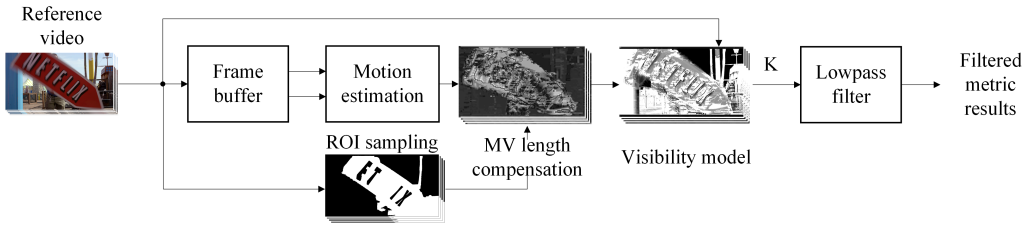


Figure 4.8: Flow diagram framework of the methodology PSNR-M+.

4.2.3 Comparison methodology

In this section, the proposed metric is compared to a range of metrics found in the literature and that are in current use. These metrics include the most popular metrics at the current moment (PSNR (Section 1.1, Page 3), SSIM (Section 2.2.2, Page 27)). Video multimethod assessment fusion (VMAF) [135], a method based on machine learning on databases of real users' evaluations of the quality of training videos [12]. Spatio-temporal reduced reference entropy difference (STRRED) [13]. STRRED utilises spatial and temporal entropic differences. The design of the STRRED algorithm was adopted using a hybrid approach that combines statistical models and perceptual principles. STRRED measures the spatial and temporal information differences between the reference and distorted videos, respectively. The objective High-Dynamic-Range video quality measure (HDR-VQM) is based on signal preprocessing, transformation, and subsequent frequency-based decomposition [136]. Video quality is computed based on a spatio-temporal analysis that relates to human eye fixation behaviour during video viewing. "FovVideoVDP: A visible difference predictor for wide field-of-view video" [4] is a video difference metric that models perception's spatial, temporal, and peripheral aspects. Metric is derived from psychophysical studies of the early visual system, which model spatiotemporal contrast sensitivity, cortical magnification and contrast masking.

The primary purpose of the analysis is to observe the behaviour of current VQM measures when deployed for predicting the perceptual quality of video content.

Table 4.2: Comparison of the video quality metrics correlation with HVS, using the Pearson correlation coefficient (PLCC).

VQA	NFLX C4S1	NFLX C4S2	NFLX C6S1	NFLX C6S2
PSNR	0.577	0.117	0.648	-0.284
SSIM	0.623	0.075	0.715	-0.314
FVVDP	0.567	0.112	0.654	-0.284
VMAF	0.594	0.169	0.568	0.164
STRRED	-0.248	-0.029	-0.451	-0.232
HDR-VQM	0.562	0.093	0.537	-0.178
PSNR-M+	0.613	0.803	0.607	0.572

The LIVE-NFLX [137, 138] data sets were used to test and compare the video quality metrics. The LIVE-NFLX data set consists of 112 compressed (hence distorted) videos, of which 12 (publicly available) are used here for testing. The LIVE-NFLX database was selected because it represented highly realistic content with the quality of experience responses to various design dimensions, including varying compression rates in the form of simulated varying transmission video bit rates throughout each video.

As seen from Table 4.2, the proposed metric gives the most consistent positive correlation. The gain in performance is mostly due to PSNR-M+'s ability to generalize predictions across video sequences over time. The strong compression distortion represented in this dataset and the difficulty of replicating subjective scores are apparent in the correlations herein. For example, HDR-VQM predicts a comparable correlation between the prediction and real subjective score estimates for only half of the studied video sequences. HDR-VQM does a good job of predicting spatial processing but does not model temporal processing well.

It is appropriate to compare PSNR-M+, in which the model of HVS is derived from data based on modern screens to FovVideoVDP [4], which was based

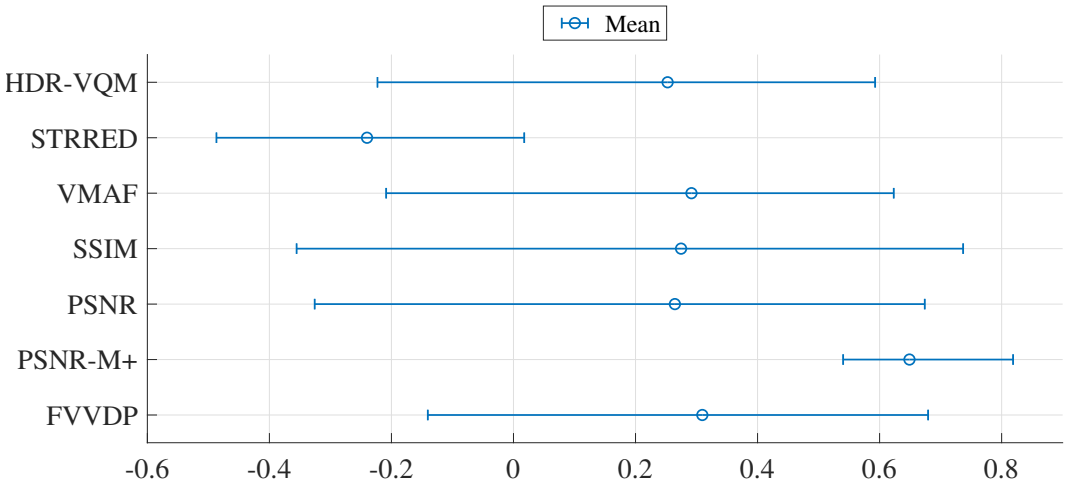


Figure 4.9: Correlation interval of video quality metrics on video sequences LIVE-NFLX. The new proposed metric, PSNR-M+, has the most consistent high correlation of the metrics tested herein.

on much older data derived from experiments using cathode ray tube screens. Both metrics consist of CSF models. FovVideoVDP shows variable results in low-contrast objects and low bitrate. When comparing FovVideoVDP with PSNR-M+, the metrics show approximately similar results for video sequences with a normal bitrate. However, PSNR-M+ gives a 15% higher average correlation for videos containing sufficient motion.

Fig. 4.9 compares the correlation intervals of video quality metrics with the real subjective score on video sequences. PSNR-M+ has a better correlation interval than the VQM algorithms under comparison. In other words, the metric more effectively predicts the perception of videos, with different content and distortions, by the human visual system.

Despite the objectively better stability of the developed method, there is still potential room for improvement. For example, the motion compensation is coarse and could be refined to reflect the motion of individual moving objects better. The size of objects selected by the ROI stage could also be refined. The use of more videos from more disparate data sets will also enhance our confidence in our correlation scores. Moreover, the HVS model is limited and does not simulate certain aspects of vision, such as inter-channel masking, eye

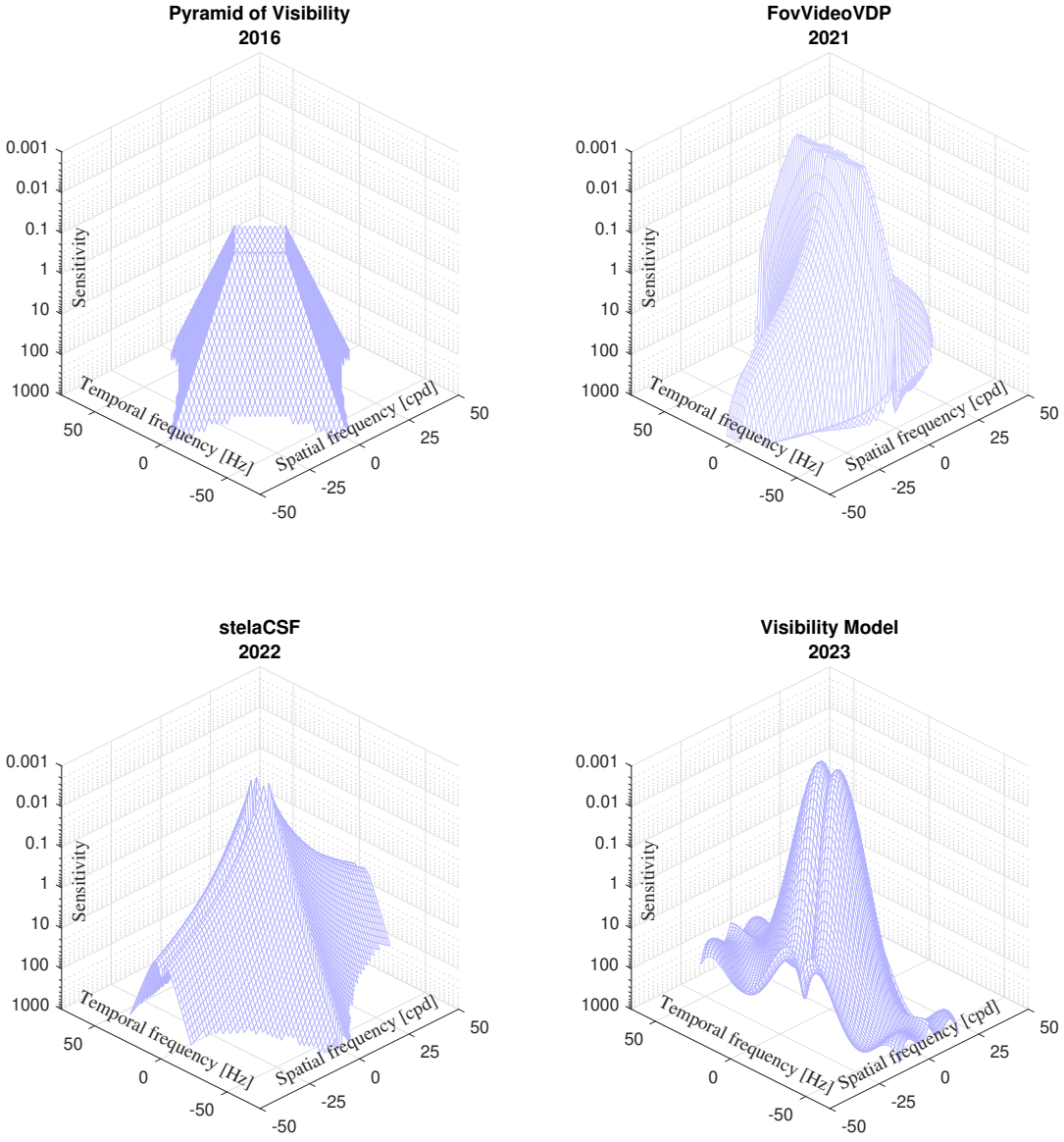


Figure 4.10: Visualization of contrast sensitivity models: Pyramida of visibility, FovVideoVDP, stelaCSF, Visibility model.

movement, and peripheral vision.

4.3 Model comparison

Comparison of existing CSF models and the model presented in this paper on existing older CSF datasets such as StelaCSF by best-fit estimation is impossible. The data on the temporal aspect in modern viewing conditions have critical differences from the older ones.

At the time of writing, two models of the CSF generated from older research

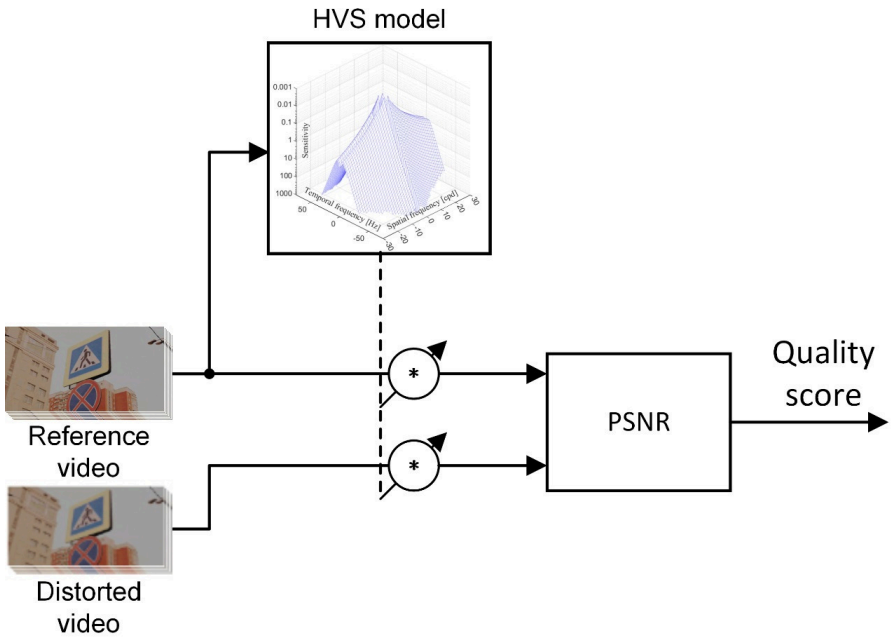


Figure 4.11: Block diagram PSNR-M+ inclusion of stelaCSF model [91], where $*$ is the convolution.

exist and consider all the basic aspects of human perception that can improve the performance of video quality assessments. The contrast sensitivity model used in the FovVideoVDP VQM [4], described above, and stelaCSF, introduced in 2022. Data from 11 early studies were pooled to create stelaCSF [91]. The stelaCSF approach can predict data from all open early studies using the same set of parameters and has accurate fit predictions over the entire range, including low frequencies. The results show that stelaCSF can explain datasets better than existing models, regardless of the number of dimensions considered, including the pyramid of visibility [91], Fig. 4.10. The model presented in the paper is a refined visibility pyramid for evaluating video content in modern conditions.

This chapter analyses two VQMs [60, 4], with different visibility models included in both metrics. In the first VQM, the inclusion of CSF models is based on calculating the weighted PSNR, Fig. 4.11.

All three models (FovVideoVDP, stelaCSF, and the proposed model) generate time coefficients that change with the frame, which are then passed to the

VQM to calculate the score value. In all models, the background brightness was set at 120 cd/m^2. The resulting contrast values following the authors [91] for the stelaCSF and FovVideoVDP models were converted from Weber to logarithmic. The model parameters were set - temporal frequency, spatial frequency and background brightness. The following databases were used to test visibility models: LIVE Netflix QoE database and MCL-V.

The comparison is not made for all parameters in the stelaCSF model. This is due to the peculiarity of the comparison since the main goal of the proposed work is to demonstrate the advantages of the obtained data on temporal contrast sensitivity in modern conditions for the provision of media content.

4.3.1 MCL Dataset

MCL-V contains 12 original video clips and 96 distorted video clips with subjective evaluation scores [139], Fig. 4.12. In the MCL-V database, the average value per video represents the subjective rating. The average value of the VQM is found for each video, and the correlation between the average values of the VQM and the average values of the subjective rating is calculated. The overall performance for all models is shown in Table 4.3.

4.3.2 Netflix Dataset

The LIVE Netflix QoE database consists of 112 distorted videos created from 14 pieces of footage at 1080p at 24, 25 and 30 fps by overlaying eight different playback templates [137, 138], Fig. 4.13. For the LIVE Netflix QoE database, the correlation value is calculated between frame-by-frame metric values and frame-by-frame subjective evaluation values. Each sequence is 10 seconds long. A video sequence longer than 10 seconds is needed for the observer to make a representative judgement, as in the real conditions of video content perception [28]. The overall performance for all models is shown in Table 4.4.

Table 4.3: Comparison of the video quality metrics on the MCL-V database. Pearson correlation coefficient (PLCC). All values represent the mean with a statistical deviation within 5%.

PLCC for VQA/ MCL-V Database	Upper	PLCC	Lower
PSNR	0.307	0.110	-0.292
PSNR-M+ with stelaCSF	0.628	0.481	0.302
PSNR-M+ with FovVideoVDP	0.64	0.497	0.321
PSNR-M+ with Visibility model	0.643	0.502	0.36
FovVideoVDP with stelaCSF	0.339	0.174	-0.038
FovVideoVDP with FovVideoVDP	0.304	0.100	-0.113
FovVideoVDP with Visibility model	0.29	0.121	-0.054

Table 4.4: Comparison of the video quality metrics on the LIVE Netflix Database. Pearson correlation coefficient (PLCC). All values represent the mean with a statistical deviation within 5%.

PLCC for VQA/ LIVE Netflix Database	Upper	PLCC	Lower
PSNR	0.34	0.318	0.297
PSNR-M+ with stelaCSF	0.384	0.364	0.352
PSNR-M+ with FovVideoVDP	0.402	0.383	0.368
PSNR-M+ with Visibility model	0.422	0.401	0.381
FovVideoVDP with stelaCSF	0.448	0.429	0.426
FovVideoVDP with FovVideoVDP	0.342	0.318	0.292
FovVideoVDP with Visibility model	0.49	0.466	0.44

Figure 4.12: Sample video frames from the Video Content MCL Database [139]. A database contains videos with artificial distortions and genre types, such as animation and sports, commonly seen in applications. For video semantics, authors consider factors that will greatly impact human visual perception.

Figure 4.13: Sample video frames from the LIVE Netflix Video Quality of Experience Database [137, 138]. A database containing videos with artificial distortions and genre types, such as rain and fast movement. This specific content is difficult to encode.

4.3.3 CSQ database

CSQ database is a large-scale set of encoded videos with constant subjective assessment, Fig. 4.14. The database uses H.264 compression, and the number of artefacts is 44800 [18]. The content in the CSQ database is a series of videos joined together to form a consistent quality sequence. For each of these videos, the average value of the VQM and the average value of the subjective opinion is found, after which the correlation between the average values is found. The overall performance for all models is shown in Table 4.5.

Figure 4.14: Sample video frames from the CSQ database containing videos with artificial distortions. For video semantics, authors consider factors that will represent as broadly as possible the types of content that might be seen in a typical streaming platform.

4.4 Summary

The presented CSF visibility model demonstrated a distinct advantage in predicting subjective video quality by testing video quality metrics incorporating the new model and other visibility models. The test was on three publicly available video datasets. Of the models tested, the VQM with the new unique model was found to have the best and comparable results in predicting subjec-

Table 4.5: Comparison of the video quality metrics on the CSQ database. Pearson correlation coefficient (PLCC). All values represent the mean with a statistical deviation within 5%.

PLCC for VQA/ CSQ database	Upper	PLCC	Lower
PSNR	0.453	0.388	0.317
PSNR-M+ with stelaCSF	0.42	0.404	0.393
PSNR-M+ with FovVideoVDP	0.381	0.363	0.343
PSNR-M+ with Visibility model	0.458	0.437	0.426
FovVideoVDP with stelaCSF	0.397	0.381	0.375
FovVideoVDP with FovVideoVDP	0.397	0.382	0.376
FovVideoVDP with Visibility model	0.34	0.339	0.328

tive video quality. The presented model has a clear advantage in terms of the scale of the data required at this stage of technology development to create VQMs based on machine learning.

StelaCSF is a model consisting of a wide variety of data today. It would be useful to add the data presented in this chapter to stelaCSF training sets. New large-scale HVS data acquired using modern screens in a slightly less controlled environment than traditionally used, can complement the currently existing stelaCSF, which is artificially created from old data.

Chapter 5

Construction of a suitable Video Quality Database

While a wide range of video quality datasets has been released in recent years [14], current databases contain a small number of video sequences with little content diversity and distortion complexity, offering a limited opportunity for developing and evaluating non-reference video quality assessments with efficient use of deep learning [15]. Large amounts of data provide deep neural networks with enough variety of examples to train on. There is ample publicly available diverse video content; however, the vast majority is not data with useful quality rates [140]. There are also a number of disadvantages to current methodologies of collecting data from subjective video quality.

First, many existing databases with quality ratings have a relatively small number of participants' quality evaluations (500–1500 per data set), which is insufficient to train deep models. The basis of this problem is the prohibitive length of time required to collect subjective assessments [141]. Until now, no optimal solution has been developed for collecting sufficient subjective assessments, which is this study's most valuable aspect. When involving participants in testing, it is necessary to create a way to conduct tests that consider the maximum possible amount of artefact processing with a minimum number of experiments [14, 141]. Doing this will help reduce data collection time, reduc-

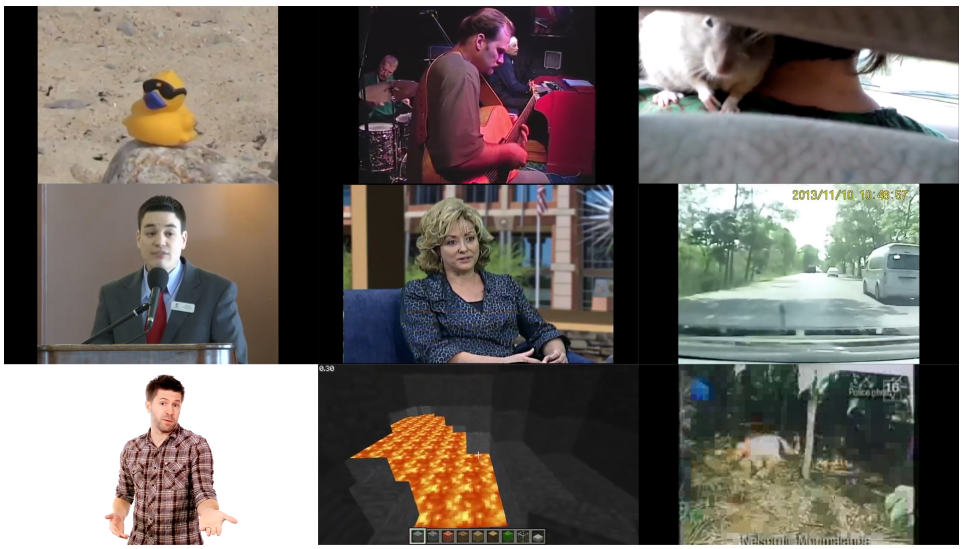


Figure 5.1: Sample video frames from the LSVQ database [142]. A database containing real-world distorted video types, such as such as nature and humans, is commonly seen in social media. For video semantics, authors consider user-generated content.

ing the financial cost of subjective tests.

Second, available databases usually include professional and laboratory videos (not videos In-the-wild). These are not representative of the user's general exposed video captures. Also, if a user provides a low rating for a particular frame, it is impossible to guess what the content should look like for a good rating, and if a user offers too high a rating, this can lead to a data set overrating [14].

Finally, there are databases containing videos shorter than 10 seconds, which does not comply with the International Telecommunication Union's video sequence length recommendations [28]. These databases should not be used when creating metrics based on machine learning. When setting subjective ratings for a video sequence, the human visual system needs time to adapt to the changing scene during a single viewing.

Currently, there is not enough well-labelled data to train and test the performance of deep learning models in video processing. There is a need for a new methodology to collect users' quality ratings, where participants can

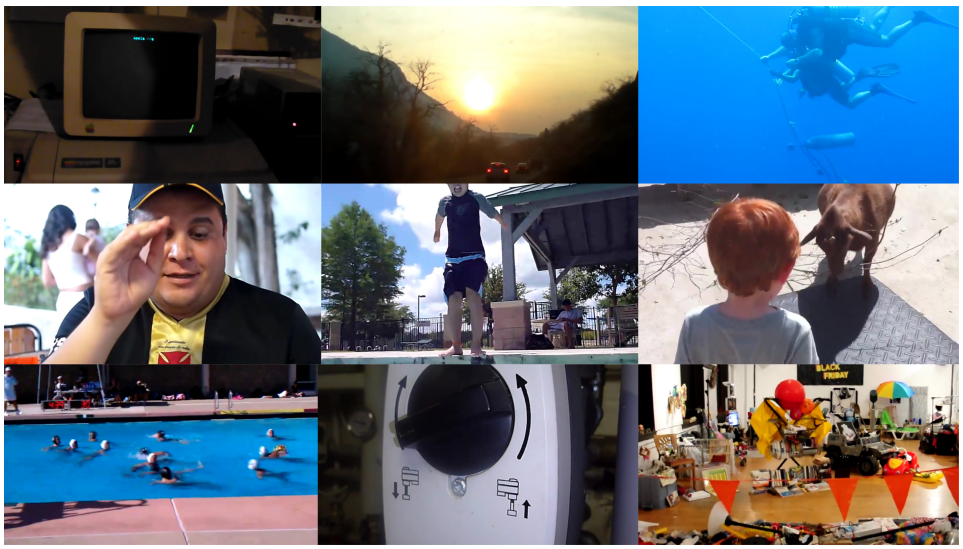


Figure 5.2: Sample video frames from the KoNViD-1k database [15]. This is a database containing natural, real-world video sequences. The video sequences are authentic 'in the wild' distortions depicting various content.

create a suitable dataset for deep learning. Here, the methodology for collecting numerical ratings or comparative analysis when creating a video database was a methodology for determining the minimum comfortable threshold of perception acceptable to the participant. The new approach is much easier and more convenient to measure. This saves time and finance on creating a database, allowing the collection of 2500 annotations in 1 hour, in contrast to the most popular data sets today, LSVQ [142] (Fig. 5.1) and KoNViD-1k [15] (Fig. 5.2), where 1760 annotations were collected, and 23 annotations were collected per hour, respectively. With this collection technique, a new dataset that provides constant quality footage created using typical media content's dynamically changing visual quality is presented in this chapter [18]. Also, in this chapter, analysis of popular video quality databases, including KoNViD-1k [15], LIVE-Netflex [135, 138], MCL-JCV [139], LSVQ [142], VQEG HDTV 7 [143, 144], and LIVE VQD [153, 154] is represented in Table 5.1.

The work in this Chapter was presented in [18, 16] during this doctoral research project.

Table 5.1: Summary of popular public-domain video quality datasets.

Database	KoNViD-1k	LSVQ	MCL-JCV	LIVE Netflix	VQEG HDTV	LIVE VQD	Proposed database (CSQ)
Different contents	1200	39075	30	14	740	15	340
Video duration (sec)	8	5−12	5	10	10	10	10−15
Distortions	1200	39075	51	112	740	150	3400
Subjective study framework	Crowd-sourced	Crowd-sourced	In-lab	In-lab	In-lab	In-lab	In-lab
Database creation time (hour)	≈ 9200	≈ 98000	3250	≈ 458	≈ 1500	≈ 525	≈ 200
Annotations (thousands)	205	5500	78	34	0.7	37	500

5.1 Analysis

The last complete review of video databases was published in 2012 and contains 13 video databases [14]. Today, the number of databases has more than doubled. Table 5.2 presents a quick overview of 90% of existing video datasets as at 2023.

Table 5.2 shows datasets where videos are shorter than the 10 seconds recommended by the International Telecommunication Union [28]. Video lengths shorter than 10 seconds are not suitable for training a deep learning system:

- Poly @ NYU Packet Loss (PL) Database Small database on the impact of packet loss on H.264 video. Test clips are only 2 seconds long [145].

- AVT-PNATS-UHD (2019) 16 UHD video sequences encoded with H.264, HEVC and VP9 and frame rate variations estimated by 29 observers [146].

- BVI-HD Perceptual Video Quality Database (2018) 22 unique HD video sequences with frame rates up to 120 Hz, assessed by 51 observers [147].

- Konstanz Natural Video Database (KoNViD-1k) (2017) 1200 videos with subjective data and attribute evaluation [15].

- LIVE YouTube High Frame Rate (LIVE-YT-HFR) Database (2020). Consists of 480 videos at 6 different frame rates from 16 different footage. Videos are processed with 5 compression levels at each frame rate [148, 149].

- MCL-JCV Database (2016) 24 original clips compressed using H.264 / AVC at quality factors (QF) ranging from 1 to 51. Subtle difference data (JND) from approximately 50 volunteers is available [139, 150].

- LSVQ Database, a subjective video quality dataset containing 39,000 distorted real-world videos, 117,000 localized spatio-temporal video patches and 5.5 million human perceptions, 38811 were used for the base with 35 estimates per video sequence [142, 151].

- VideoSet (2017) The database includes 3520 sequences, which 800 participants evaluated [152].

As shown in Table 5.2. the following databases have a relatively small number of participants' quality evaluations, which are not suitable for training a deepleaning system:

- LIVE Video Quality Database. Test conditions include MPEG-2 compression, H.264 compression, and the simulated transmission of compressed H.264 bitstreams over error-prone IP wired and wireless networks [153, 155].

- ETRI-LIVE Space-Time Subsampled Video Quality (STSVQ) Database (2020). 437 videos were created by applying different levels of combined space-time downsampling and video compression on 15 different video content and 15,000 subjective video quality ratings [156].

- LIVE Wild Compressed Video Quality Database (2020). 55 different help videos (content) contained in the LIVE VQC database, each 10 seconds long. UGC videos are captured with various mobile cameras, covering a wide range of content and quality. Most of these videos are distorted with various authentic mixed distortions when captured. H.264 video compression format [157].

- EPFL / PoliMI Video Quality Assessment Database. Testing conditions focused on compressed H.264 videos corrupted by simulated packet loss due to transmission over an error-prone network [158, 159].

- IRCCyN / IVC 1080i Database contains high-definition (HD) video compressed using H.264. In addition to ACR MOS, SAMVIQ MOS is available for part of the database [160, 161].

- IRCCyN / IVC SD RoI Database includes standard definition (SD) video compressed using H.264, with and without transmission errors [162, 163].

- IVP Database contains progressive HD video compressed using MPEG-2, Dirac wavelet and H.264 codecs, as well as H.264 streams that are affected by packet loss simulation. DMOs are provided separately for expert and non-expert observers [164].

- MMSP 3D Video Quality Assessment Database is the first publicly available 3D video quality database. Test conditions represent different distances between cameras. All videos are slightly cropped and compressed [165, 166].

- MMSP Scalable Video Database (SVD). Test conditions include two scalable video codecs using different spatial and temporal resolutions. The

database only includes the sources, the software, and the test creation process and does not include test videos per se. The subjects made paired comparisons in parallel viewing sessions [167, 168].

- Poly @ NYU Video Quality Databases. Three separate but related tests using video with different frame rates and quantization parameters [169, 170, 171, 172].

- VQEG FR-TV Phase I Database. The oldest publicly available quality database (interestingly, it started a few years before the first image quality database). Consequently, the test environment focuses on MPEG-2 compression and transmission, including some analogue distortion [173, 174].

- VQEG HDTV Database. Test conditions include MPEG-2 and H.264 compression and various types of network disturbances. Five of the six kits in the HDTV test are released through the Consumer Digital Video Library (CDVL); the sixth set is not publicly available [35, 175].

- BVI-HFR High Frame Rate Video Database (2015) 32 reference and 384 distorted sequences generated using both original High-Efficiency Video Coding (HEVC) and HEVC with synthesis mode, evaluated by 86 observers [176].

- LIVE Mobile Video Quality Database (2012) Test conditions focused on compressed H.264 video with artefacts such as packet loss, frame freeze, and rate adaptation [177, 178, 179].

- LIVE Netflix Video Quality of Experience Database (2017) 112 videos of typical adaptive streaming artefacts rated by 55+ people on a mobile device, Fig. 4.13 [137, 138].

- LIVE Video Quality Challenge (VQC) Database (2018) 585 videos captured using 101 different devices with a wide range of complex, reliable

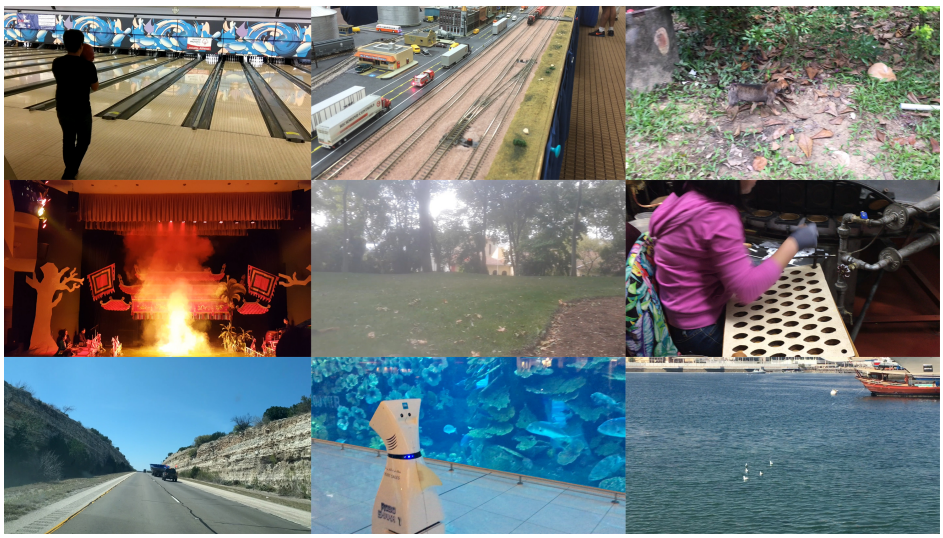


Figure 5.3: Frames from the LIVE Video Quality Challenge (VQC) Database [180, 181, 182]. A database containing videos with artificial distortions and genre types, such as nature and urbanization. This specific content is difficult to encode.

distortion levels. An average of 240 quality ratings was collected for each video through crowdsourcing, Fig. 5.3 [180, 181, 182].

- TUM 1080p25 Dataset (2010) 1080p25 SVT test suite video sequences, encoded with H.264 / AVC and Dirac [183].

- LIVE-NFLX-II The database includes 420 videos rated by 65 subjects, resulting in 9750 continuous and 9750 retrospective subjective opinions [2].

All presented datasets have flaws that are critical for testing deep learning models. Another difficulty when comparing existing databases is when authors provide incomplete information on the database creation (shown as empty cells in Table 5.2). Also, existing popular databases have very different and often almost incomparable approaches to the implementation of the collection of subjective data.

Existing databases can be divided into several types of subjective testing. LIVE [155], LIVE-YT-HFR [148], LIVE-NFLX-II [2], ETRI-LIVE STSVQ [156],

and LIVE Wild [157] have a continuous quality assessment protocol [28]. Consider the how the example of LIVE Wild [157]. The videos in LIVE Wild were played randomly, each being shown only once during each session, and at least five videos were split between different distorted versions of each unique content. The adjectives Bad and Excellent denoted the quality range from low to high. Subjective scores received from participants were converted to numerical quality scores in [0, 100].

A different method to that above was used to create databases AVT-PNATS-UHD [135], MCL-JCV [184] and VideoSet [152], each of which contains, according to the authors, a large number of processed video sequences with different levels of processing. However, of the 30 clips declared in MCL-JCV, only 24 sequences were publicly released due to an intellectual property problem [150], and the clips are only five seconds long. It is also worth noting that these databases' distortions were created synthetically. The subjects compared the quality of the two sequences displayed one after the other and determined whether the two sequences were significantly different by choosing yes or no. This approach provides more data than other existing methods of creating databases for estimating video by people. However, MCL-JCV and VideoSet databases do not provide enough video data for deep learning (only 5 hours). Also, the data sets do not indicate the number of evaluations per sequence. Even with a minimum number of 10 evaluations, the ratio of the time spent on subjective tests (people's time) to the final total database time is significantly inferior to the first approach. In addition, it is not clearly understood how the threshold of the minimum acceptable quality for a person is determined in a database (when the quality ceases to suit the participant), which makes the database biased to high quality.

LIVE VQC [180] and KoNViD-1k [15] databases use the crowdsourcing method of subjective score collection. Crowdsourcing method [15] when the researcher first provides participants instructions on the types of artefacts (e.g., related to motion, colour, brightness, and detail) and how to rate the overall

quality of each video. Then, examples of videos with, for example, Good, Satisfactory, and Poor quality are shown to participants for reference. At the end, a rating scale is displayed. Only when a participant has viewed and rated all videos on the screen, it is possible to move on to the next step. Gold standard questions are commonly checked to monitor the quality of the participants' work. This approach shows significantly worse results compared to previous methods, Table 5.3.

A complete overview of the most popular video datasets is presented in Table 5.3. The table shows that a solution for one or more problems presented in this chapter for testing deep learning models in video processing, as with the VideoSet database, does not necessarily result in the best database.

In video encoding, the subjective tests show that humans perceive video quality levels in discrete steps over a wide range of encoding bitrates, i.e. as a step function of the quantization parameter [152]. The values at which a jump between quality levels occurs are called just-noticeable difference points. The literature presents an exhaustive description of the need to consider the transition of thresholds when collecting subjective tests and the largest-scale subjective study to create a database [152]. As shown above, this fact is only considered in limited data sets.

Considering the shortcomings described above and the analysis presented in Table 5.3, a new methodology for collecting the subjective rating is required. The new methodology should also address a significant problem where evaluators find it difficult or impossible to quantify the user experience of viewing media content with changes in quality. That problem was addressed in the next section.

5.2 Method for Creating CSQ Database

Reference video clips created by a RED Comodo 6k camera with Sigma optics were chosen as the source sequence; see Fig. 4.14 and Fig. 5.4. The videos

Figure 5.4: Figure 5.4: Sample video frames from the new database containing videos with artificial distortions. This chapter considered factors in various video features for the database semantics.

in the database are stored in YUV422 (YUV is a colour space that separates brightness information from colour information in the video, YUV422 is a data format that shares U and V values between two pixels) uncompressed format with a fixed resolution of 1920 × 1080 at 24 frames per second. The average length of a video sequence is 10 minutes. The video sequences for the database were professionally recorded by the expert photographer in digital format. This method allows the distortion of the video and subjects it to several processing stages, both on the device itself and the server.

The content of the database is based on three criteria. Firstly, this is user content, unlike other databases such as LIVE [155], LIVE-NFLX-II [2], ETRI-LIVE STSVQ [156]. In the presented database, an attempt was made to view typical contemporary multimedia content scenes. Secondly, the database must have sufficient variety in several characteristics; here, we follow according to the criteria presented by Winkler for original content [14]. The presented database (CSQ) divided the original content into three groups [14] : (1) high-level video genres, (2) mid-level video semantics, and (3) low-level video features. The CSQ database covers a wide range of characteristics, as shown in Table 5.4. Third, optimization of the collection of subjective ratings is considered.

The proposed method of creating a dataset generates levels with different quality from the reference uncompressed video using H.264 compression, the most popular video format for video streaming over the network. A two-pass coding scheme was used to provide a consistent quality of perception frame by frame for a subjective real-life assessment. Input and output video resolution are supported at 1080p. An X-rite i1 Pro (X-Rite Inc., USA) device was used to test the monitor color profile (d65).

The minimum allowable distance from the monitor to the participant for finding a stimulus in the clear vision zone is 0.872 m, and the maximum is 1.149 m [17], Chapter 3 (Section 3.1, Page 53). 36 of the 40 participants were third-year undergraduate students, which is a good balance between three important parameters: physical maturity of the eyes, daily use of typical user-generated content, namely watching videos and images on the internet and lack of experience with visual information on perception. Forty participants took part in the experiment (28M, 12F, aged 18–45, normal or corrected-to-normal vision).

A block diagram of the method is shown in Fig. 5.5. First, video sequences distorted with ten quality levels were created. In modern streaming video, the most demanded levels range from 6 to 10 Mbps. However, for the accuracy of the experiment, all levels are needed, including the lowest and highest. The presented work used levels from 0.2 to 40 Mbps. The one testing video sequence consisted of 9−10 video sets of 5−6 videos 10–15 seconds, each video set with different content. The participants in the experiment found the minimum acceptable threshold perception for each video sequence, constantly adjusting the quality with the help of a manipulator and overcoming the tension of a manipulator [18], Fig. 5.6. Overcoming tension, or pressure on the pedal, ensures constant participant involvement in the experiment. Otherwise, the participant would have just set the manipulator to maximum quality. Pressure on the pedal forces the participant to constantly lower the quality (loosen the pressure on the pedal) and find the minimum acceptable level of quality when

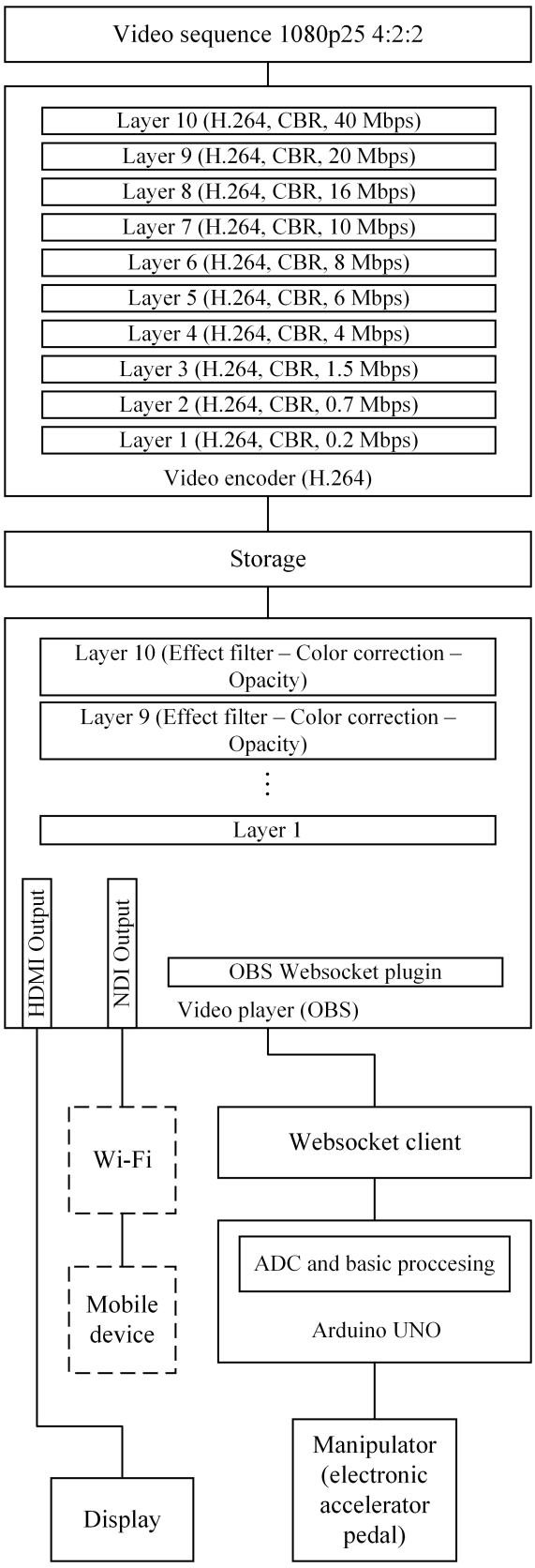


Figure 5.5: Structural diagram of the installation for research.

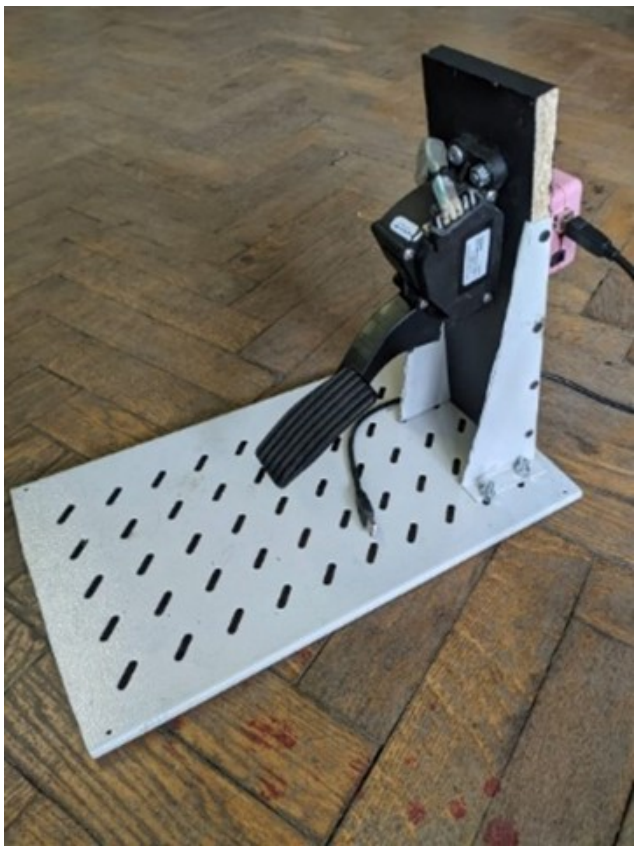


Figure 5.6: The manipulator.

artefacts are not noticeable. As a manipulator (pedal), the electronic gas pedal of the Lada Priora car, which has satisfactory ergonomics, was used.

The present work's perceptual threshold is the acceptable minimum level of video quality for a participant at which video viewing is comfortable. Before the tests, there were reasons to believe that the acceptable level of perception could be very different for each user; however, it was not found to be the case in practice, similar to other psychovisual experiments [185]. The subjective quality of the content changed from time to time, and the participants were allowed to adjust the quality level using the manipulator. The length of one video sequence is from 6 to 8 minutes, and the length of the video inside the sequence is 10−15 seconds. A grey background separates all videos within the sequence. When viewing a grey background between videos, participants set the manipulator to a position of satisfaction according to instructions. At the end of each session, scores were automatically collected anonymously and

recorded in a spreadsheet. The experiment finished when the experimental uncertainty, as measured by the confidence interval, became less than 5% of the current value for all tests performed during the experiment. The HECS Human Ethics Committee approved the study at the University of Waikato (HECS−22−01).

5.3 Testing Methodology

For fair methodology testing, we follow the following three steps. First, the average spread of values for each frame will be determined. This step should demonstrate the correct completion of the experiment. A 95% confidence interval will be used for the data presented. All experiments were to continue until the confidence interval fell below 5% of the current mean for each frame across the participants. In other words, the experiment must be repeated until the reliability criterion is met. The standard deviation σ_{kfl} for confidence interval estimation for each representation is given by International Telecommunications Union [28].

The second testing step will involve comparing a high-quality sequence with the experimentally obtained sequence of an acceptable minimum threshold. The high-quality sequence will be the reference recorded video with H.264 compression (40 Mbit) comparable to level 10 in Fig. 5.5. This analysis should demonstrate whether the user's perception of the minimum acceptable level matches the perception of high-quality video. With a positive result and a high correlation, it is possible to assert that this methodology does not give underestimated results (quality not acceptable to the user). As described above, the problem of overestimating quality was solved using pressure in the manipulator. A full-reference quality metric- the maximum signal-to-noise ratio PSNR will be used to compare video sequences. PSNR is the most common VQM in the world [9]—almost all video quality studies that may use full-reference values use PSNR along with others because PSNR based

on the standard deviation is never overestimated. For analysis, PSNR values will be calculated for the experimentally obtained video sequence of constant subjective quality and the high-quality video sequence described above.

The third step for analyzing the proposed methodology is to compare presented subjective assessments with the subjective assessment obtained according to the method used to acquire the LIVE-NFLX database [137]. The data collection methodology presented in LIVE-NFLX is one of the most popular today. For the analysis, subjective assessments will be used from the LIVE-NFLX database (reference video number 4 with H.264 compression; see Chapter 4, Section 4.3.2, Page 82). Compressed video sequences with ten different quality levels were created from the same reference video to perform this third step. This analysis will determine whether data collection participants using the presented methodology can reliably provide ratings based on the quality of the video displayed.

5.4 Results

The obtained subjective data were processed, and the average rating of quality level for each frame and the confidence interval were found. On average, the scatter of values of participants' rating of quality level is 3% across participants for each frame. The example of the spread of results is shown in Fig. 5.7. Based on the average values, sequences with a constant subjective score were generated.

Fig. 5.8 shows PSNR values for the high-quality sequence and PSNR for an experimentally obtained sequence of an acceptable minimum threshold. The correlation between video sequences is 0.97. The comparison of video quality assessment metrics (PSNR) confirms that the presented methodology does not give underestimated results (quality not acceptable to the user).

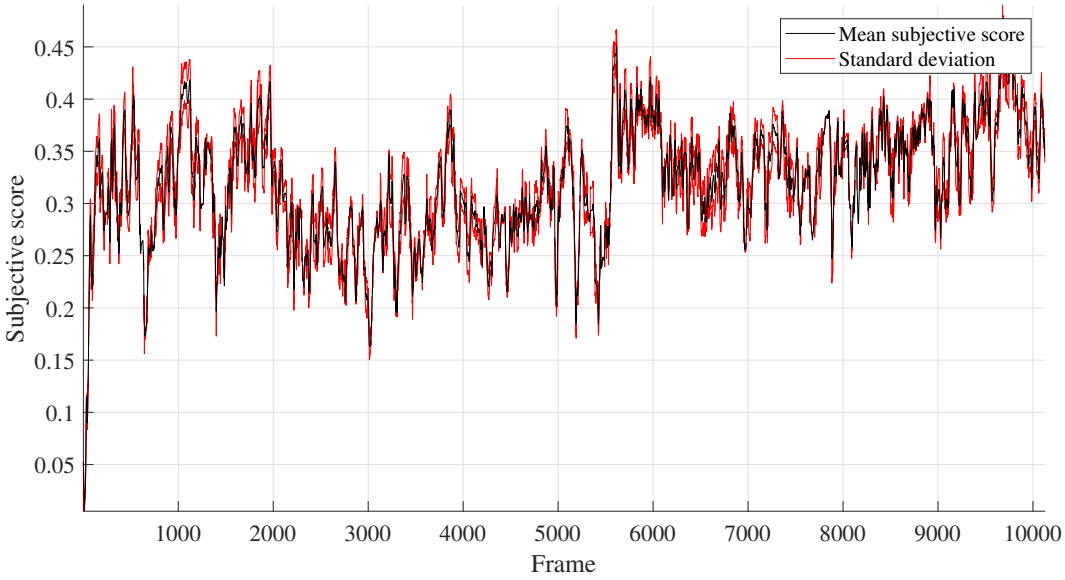


Figure 5.7: The average rating for each frame and the confidence interval across participants.

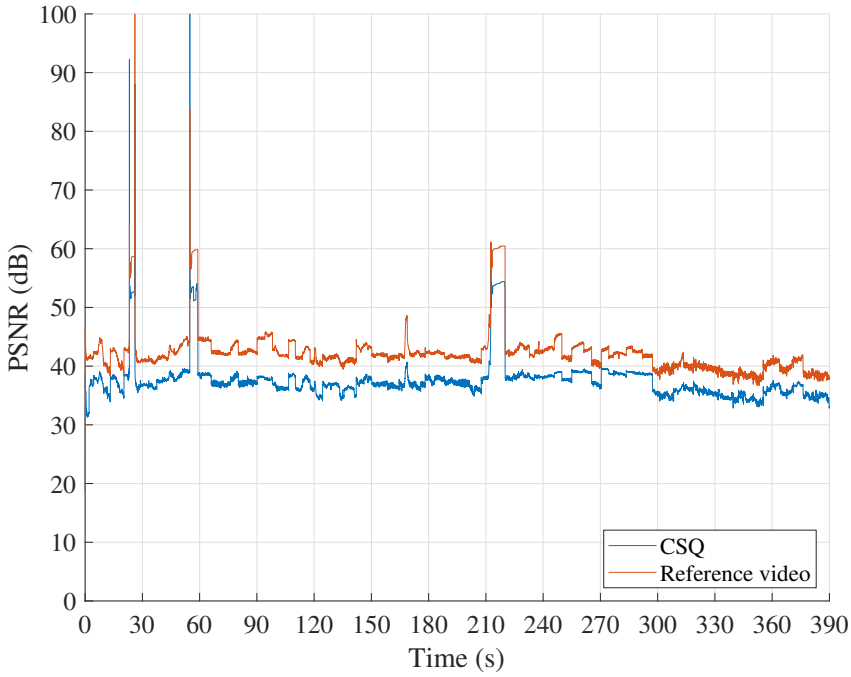


Figure 5.8: Comparison of PSNR values for the high-quality sequence and PSNR for an experimentally obtained sequence of an acceptable minimum threshold.

5.5 Summary

In this chapter, a new device for measuring the quality of the encoded video was created based on finding an acceptable minimum perception threshold,

which allows for generating video sequences of constant quality. The new CSQ database was also presented, which contains 340 unique contents and close to 500000 annotations. The database is publicly available for future research and development, https://github.com/anastasiamozz/CSQ-Database.

At the current stage of technology development, there are more than seven methods for collecting subjective assessments of video quality and more than 27 databases in the public domain. This makes it easier to test video quality metrics, but still not enough to create new quality assessment algorithms or video codecs based on full or partial machine learning. The approach proposed in this chapter allows obtaining video with a constant quality assessment created by the users themselves, which codec developers should strive for as optimal for users. CSQ method and device will allow the creation of new databases with different distortions and well-labelled data in a larger volume than the currently used. A specialized installation that dynamically changes the visual quality of typical media content during the research process can be used to create effective adaptive codecs built on a hybrid model and for new ones based on machine learning.

Table 5.2: A quick overview of existing video datasets in 2023. The incomplete information on the database creation is shown as empty cells.

Database	Number of sequences	Number of artefacts	Number of participants	Length of one sequence (second)	Number of experts for 1 sequence
EPFL/PoliMI -1	156	78	40	10	23
EPFL/PoliMI -2	156	78		10	17
IRCCyN/IVC 1080i	216	192	29	10	28
IRCCyN/IVC SD	90	84	25	10	25
IVP	138	128	42	10	35
LIVE VQD	165	150	38	10	38
MMSP 3D	36	30	20	10	17
MMSP	87	84	16	10	16
Poly@NYU(VQD -1)	66	60		10	22
Poly@NYU(VQD -2)	72	68		10	15
Poly@NYU(VQD -3)	186	180		10	15
Poly@NYU (PL)	46	34	32	2	32
VQEG FR-TV	340	320	287	10	61
VQEG HDTV	740	740	120	10	24
AVT-PNATS-UHD	4947	4947	121	7	24
BVI-HD (VQD)	416	384	86	5	86
BVI-HFR	110	88	29	10	29
KoNViD-1k	1200	1200	642	8	114
LIVE-YT-HFR	496	480		8	40

Database	Number of sequences	Number of artefacts	Number of participants	Length of one sequence (second)	Number of experts for 1 sequence
LIVE Wild	275	220		10	40
LIVE mobile (VQD)	210	200	38	15	27
ETRI-LIVE-STSVQ	452	437		10	34
LIVE Netflix	126	112	55	10	44
LIVE (VQC)	585	585	4776	10	240
MCL-JCV	96	1124	120	5	50
TUM 1080p25	52	48	19	10	19
VideoSet	3520	44880	800	5	
LSVQ Database	38811	38811		7	35
LIVE-NFLX-II	435	420	65	10	22

Table 5.3: A complete overview of the most popular video datasets in 2023.

	Database name					
	MCL-JCV	Live V	VideoSet	LIVE Wild	CSQ	KoNViD-1k
Subjective Assessment Procedure	Just noticeable difference	Single-stimulus continuous quality assessment protocol	Just noticeable difference	Single-stimulus continuous quality assessment protocol	Threshold search	Crowd-sourcing
Number of reference videos	24 to 5 sec	10 to 8.68−10 sec	880 to 5 sec	55 to 10 sec	340 to 10−15 sec	not applicable
Number of artefacts	1224	150	44880	220	3400	1200
Number of dataset's videos	96	160	3520	275	340	1200
Number of participants	120	38	800	40	40	642
Number of videos analyzed by one participant	10	160	714−765	140−135	1700	550
Time for one test, (min)	40	30	35	45	25	not public information
Number of evaluations	50	38	not public information	40	20	114
Total test time and time's length of database	80 hours and 6 minutes	19 hours and 26 minutes	1000 hours and 5 hours	60 hours and 46 minutes	8 hours and 5.5 hours	380 hours and 3.5 minutes
Time spent on processing one video, (min)	50	7.6	45	16	3.5	19
Time of subjective tests by all participants	80 hours− 612 videos	19 hours− 60 videos	1000 hours− 224400 videos	60 hours− 275 videos	8 hours− 340 videos	380 hours− 1200 videos

Table 5.4: Variety of video features.

Video genres	Video content	Changing video characteristics
On the street	Persons	Brightness
Indoors	Faces	Contrast
	Snow	Texture
	Panorama	Movement
	Objects	Color Variation
	Artificial objects	Camera movement
		Scene change

Chapter 6

New Real-Time Video Quality Metric

There is a strong need for non-reference video quality metrics for user-generated video content to prevent loss of video quality caused by distortion during recording, compression, and signal transmission (Chapter 1, Section 1.1, Page 5). This chapter discusses deep learning-based non-reference video quality metrics because the ability of deep learning to model complex patterns and adapt to diverse scenarios allows algorithms to achieve better alignment with human perception [25]. Also, the need for non-reference video quality metrics based on deep learning is driven by several key factors related to the limitations of traditional approaches and the demands of modern video systems: practical applicability in real-world scenarios, the increasing complexity of video systems, limitations of traditional metrics, and advances in deep learning. For example, the reference video may not always be available in real-world scenarios, especially in streaming and broadcasting. Also, adaptive bitrate streaming dynamically adjusts resolution and bitrate to match network conditions, and the reference video might not always match the format of the distorted video, making using full reference video quality metrics impossible. It is also important to consider that new technologies like high dynamic range, wide colour gamut, AI-based systems, and high frame rate add more variables

to the quality equation, which full-reference metrics may not be designed to handle. Chapter 6 contributes to advancing the issue of streaming quality by creating a new first non-reference video quality metric that includes the psychophysical features of the user's video experience, which provides stability in predicting the user's subjective rating of a video. The experimental results show that the proposed video quality metric achieves the most performance stability on three independent video datasets.

The work in this Chapter was presented in [16] during this doctoral research project.

6.1 Modeling Predictor

In this chapter, the input videos were not cropped, subsampled, or processed. Unprocessed input videos in deep learning have the advantages of preserving contextual information, simplifying the data preprocessing pipeline, and increasing reproducibility. The proposed approach is to extract the spatio-temporal features on the raw source video, based on the knowledge of human visual models in the zone of clear and peripheral vision (local and global patch), see Fig. 6.1.

The process begins with input videos from the CSQ and MCL databases in the algorithm, each video labelled with subjective quality ratings. A high-dimensional color transform (HDCT) is applied to detect salient regions (local patch) within frames in the HDCT ROI block, highlighting areas that draw viewer attention [186]. HDCT is a technique developed for areas such as video compression because it maps colour data into a higher-dimensional space, often improving the separation of features or patterns in the data. Also, HDCT's enhanced feature representation aligns well with machine learning models, improving their performance in tasks such as object detection, facial recognition, and other vision-related tasks [186]. The local patch block extracts these regions using a mask whose width is 20% of the frame width and whose height is

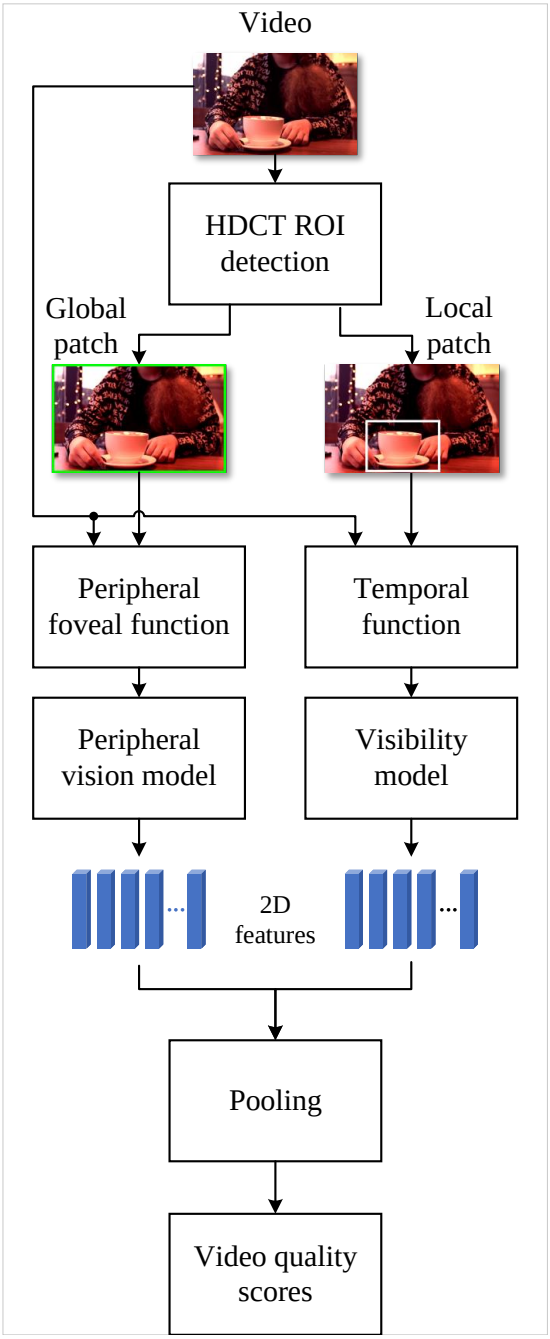


Figure 6.1: Modelling local to global perceptual quality.

20% of the frame height, while the global patch block keeps the entire frame intact. Motion vectors are computed in the foveal function block to simulate how human vision tracks dynamic elements. The peripheral model decreases the importance of frame pixels as their distance from the centre of gaze (centre of the region of interest) increases. This work used the arctangent (arctan) function because it preserves smaller gradients in areas with less variation [187]. A visibility model is then applied as a filter to retain perceptually significant

features in regions of interest. Two-dimensional features capturing spatial and temporal data are extracted in the 2D features block. Finally, a deep learning pooling method reduces spatial dimensions while preserving critical information for assessing video quality.

When watching a video, a specific part of the frame attracts the viewer's attention, and the human observer captures a particular part of the frame, called the user's region of interest or centre vision area. The region of interest or centre vision area exists because photoreceptors (cones and rods) and ganglion cells are unevenly distributed in the human retina [132]. Cones are responsible for light vision at high light levels, and their density determines visual acuity (Chapter 2, Section 2.3.1, Page 33). The highest density of cones is found in the fovea, an area near the retina's centre.

The proposed metric at the first stage determines the user's visual focus in the frame or, in other words, determines the region of interest (the zone of centre vision). The repeatability of the user's areas of interest during multiple viewing of one video sequence always has deviations of no more than 5% [185], as described in Appendix A.2. The perceived visual quality of signal artefacts depend slightly on the physical screen sizes, Appendix A.3. That allows to use the algorithmic determination of the areas of the definition of clear and peripheral HVS. Finding regions of interest is implemented by a computer algorithm that automatically detects salient regions of a frame using a multidimensional colour transform. The basic idea is to represent the frame significance map as a linear combination of a multidimensional colour space that can clearly distinguish between salient regions and the background. The highlighted areas often have distinctive colours compared to the background in human perception, but it is worth noting that human perception is often quite complex and non-linear.

For the creation of non-reference VQM used low-level visual modelling based on psychophysical models, namely the contrast sensitivity function [91]. Human visual sensitivity can be described as a reference filter in terms of

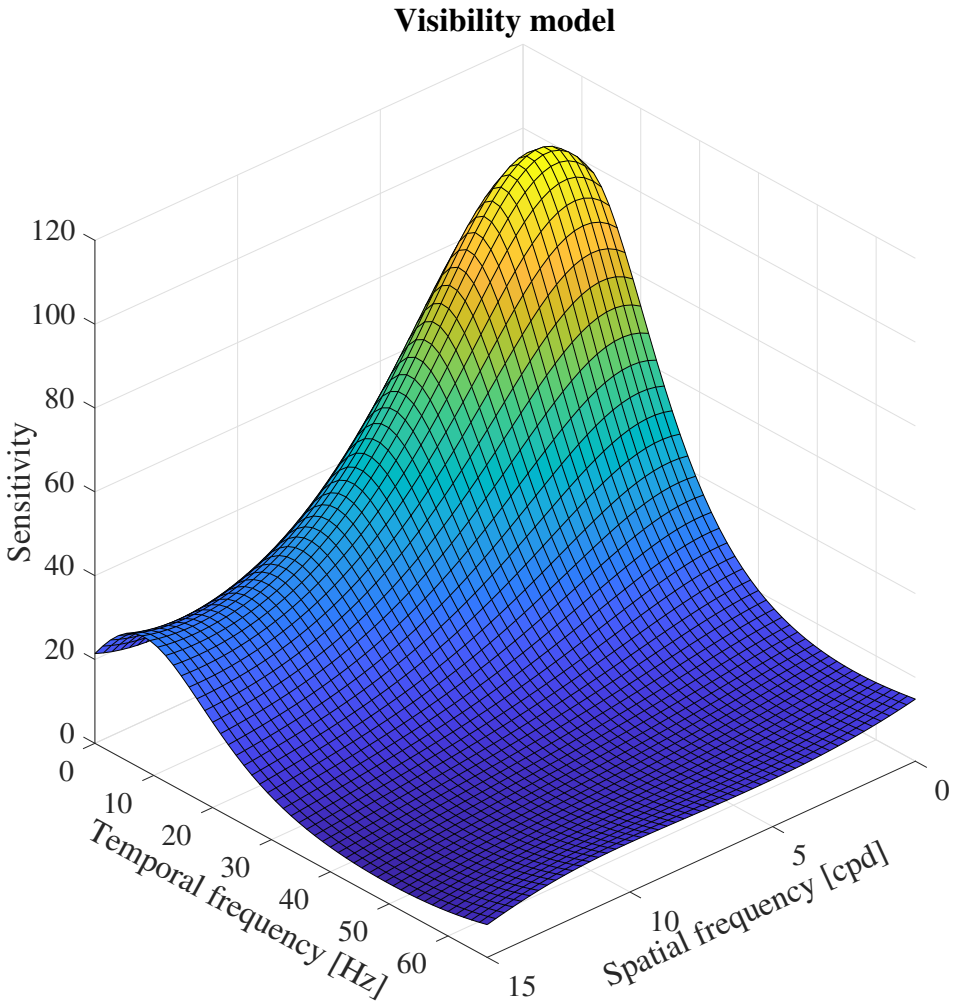


Figure 6.2: The visibility model from Chapter 4, that is, the spatial-temporal frequency response of the human visual system.

spatial and temporal frequencies, and there is no need to render outside the boundaries region [87]. We used a model for the centre vision created in Chapter 4, Fig. 6.2. The effects of peripheral vision are considered according to the spatial characteristics of the distribution of receptors [132].

Partitioned video frames are divided into global and local patches, Fig. 6.3. The local patch represents the part of the frame inside the region of interest, while the global patch represents the rest of the frame. Spatial and temporal characteristics for each frame (2D) were extracted using human visual system models. After splitting into patches, the global patch passes through the peripheral vision model (Chapter 4, Section 4.2.2, Page 73), which allows considering the change of contrast sensitivity threshold in regions of the frame

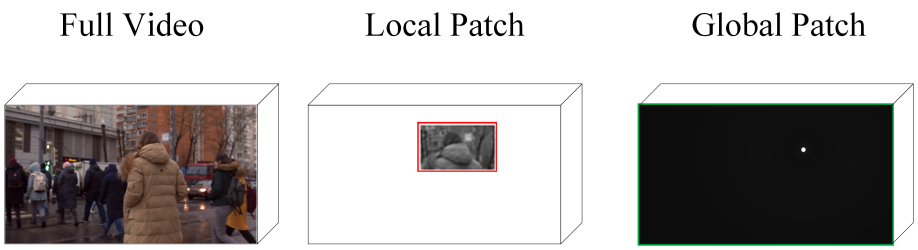


Figure 6.3: The information of video patches. A full video frame (left), a local patch filtered by the visibility model (centre) and a global patch filtered by the peripheral model (right).

depending on the distance of the region of interest from the supposed centre of gaze (in our case, from the centre of the region of interest). The weights are obtained by multiplying the global patch by a peripheral vision model that operates as a radial gradient centred on the user's region of interest. Weighting coefficients are provided to the input of the neural network. The local patch is filtered by a model of contrast sensitivity of the human visual system without considering the peripheral component, described in Chapter 4 (Section 4.1, Page 68), Fig. 6.3.

The next step of feature fusion is a process in which the features extracted by convolutional neural networks from patches processed, in one case by the visibility model and in the other by the peripheral model, are fed to the input of one fully connected neural network. After that, the neural network is trained, or in other words, the data results are considered after processing two patches to determine the value of the overall quality assessment. Training implies a process in which a labelled training data set uses the training method (gradient descent method) to change the values of neural network weights so that the prediction error tends to zero over the training time. The neural network can then predict the score on other data sets. The obtained characteristics of the quality of space-time are transferred to a modern deep model for time. The regressor consists of a series of initial modules, a global pooling average, and a fully connected layer. The initial modules study changes in quality characteristics over time, which is critical to accurately predicting

video quality.

We used the CSQ database for deep learning. The number of processed artefacts is sufficient for deep learning (a plateau was reached) and is 15000 for each video sequence of the dataset. The model was created and trained using the TensorFlow framework, and training lasted for 100 epochs. To train the model, local and global patches obtained from CSQ sequence frames from Chapter 5, Section 5.2, Page 96 were used, labelled with the number of the quality level selected by the subjects as acceptable. The new metric efficiently captures perceptual quality changes over time and predicts one quality score for each video frame, predictor architecture presented in Fig 6.4. Additional training was performed using the MCL video database, Fig. 4.12 (MCL database [88] detail information presented in Chapter 4, Section 4.3.1, Page 82) to improve model predictions in the near-threshold region. The output of the trained model can be given values in the range from -1 to 1, where 0 is the CSQ database threshold of good video quality, Fig. 6.5.

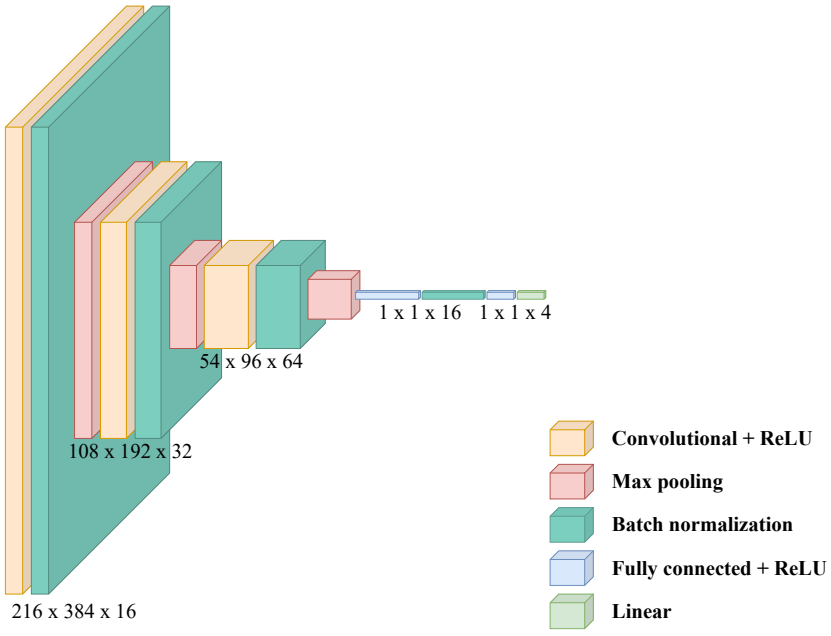


Figure 6.4: Predictor architecture for quality assessment. The predictor consists of three convolutional layers, three pooling layers, and two fully connected layers followed by a fully connected layer with linear activation.

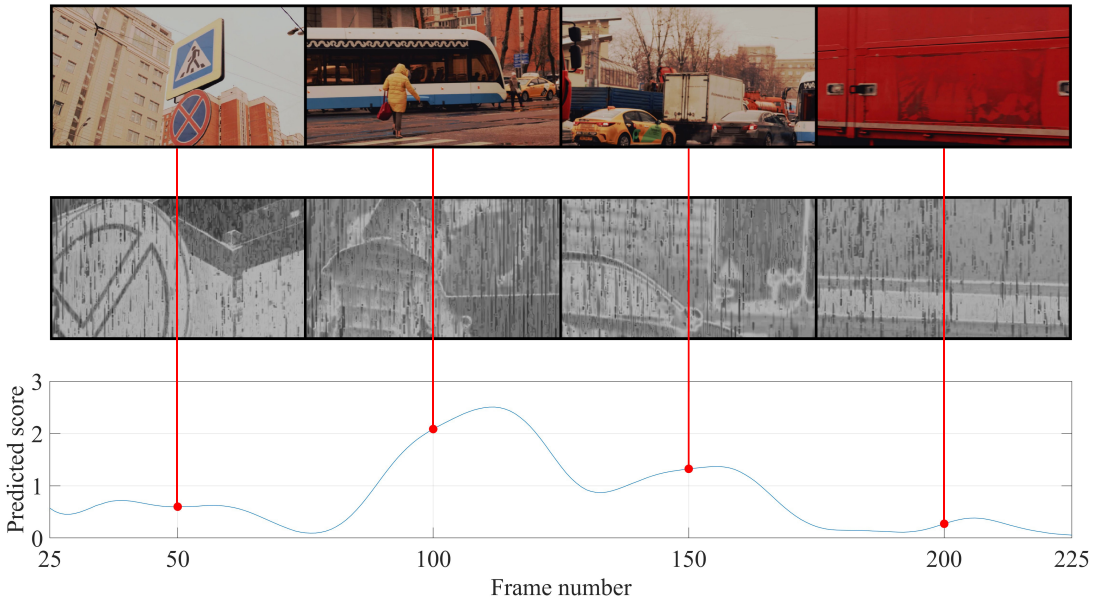


Figure 6.5: The top row is the 50th, 100th, 150th, and 200th frames. The middle row shows a local patch. In the bottom row, we give the predicted score by the proposed metric.

6.2 Metric Comparison

The presented dataset of videos was divided into training 60%, validation 20% and test set 20%. Spatio-temporal features were extracted through processing by the visibility and peripheral vision models and then marked by subjective scores. These features were passed through a pooling layer, increasing the spatial component's entropy. The pooling layer is designed to reduce the size of the coefficient grid. Thus, dividing the grid into 3x3 segments gives us a three-fold reduction on each side. In this case, entropy increases due to discarding coefficients with small values (MaxPooling). Then, the time series regressor was used, which also allowed the time component to be considered, (considering a statistical method for predicting a future response).

The proposed NR VQA named NRsptemVQA is compared with the world's most widely used full reference metrics SSIM, PSNR, full reference metrics containing models of the human visual system FovVideoVDP [4] and PSNR-M+ [60], and most popular NR video modes blind/referenceless image spatial quality evaluator (BRISQUE) [188] and VIDEoquality EVALuator (VIDE-

VAL) [189]. The BRISQUE video quality metric uses scene statistics of locally normalized luminance coefficients [188]. The algorithm only quantifies the 'naturalness' (or lack thereof) in the image due to the presence of distortions. The predictors (features) used in the training phase are derived from natural scene statistics, the locally normalized luminances computing via local mean subtraction and divisive normalization. Such an operation is applied to a given intensity image $I(i,j)$ to produce coefficients:

$$\hat{I}(i,j) = \frac{I(I,j) - \mu(i,j)}{\sigma(i,j) + C}, \tag{6.1}$$

where $i \in 1, 2...M, j \in 1, 2...N$ are spatial indices, M, N are the image height and width, respectively, and $C = 1$ is a constant that prevents instabilities from occurring when the denominator tends to zero. After this, from the coefficients, the parameters of the generalized Gaussian distribution are estimated [188], where the generalized Gaussian distribution with zero means is given by:

$$f(x;\alpha,\sigma) = \frac{\alpha}{2\beta\Gamma(1/\alpha)} \exp\left(-\left(\frac{|x|}{\beta}\right)^{\alpha}\right), \tag{6.2}$$

where

$$\beta = \sigma\sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}}, \tag{6.3}$$

where Γ is the gamma function. Next, a regression model (usually a support vector machine) is trained on these parameters, which are combined into a 36-dimensional feature vector from a dataset of images with known subjective quality scores [188]. The regression model predicts an image's perceived image quality score, a numerical value, with lower scores indicating higher quality. VIDEVAL uses a feature ensemble and selection procedure on top of existing efficient VQA models. VIDEVAL effectively balances the trade-off between VQA performance and efficiency [189]. The Patch-VQ metric for deep learning [142] is not open source and hence could not be compared. Eleven models were trained and evaluated on the same train/test splits, and as common practice, in video quality scoring, performance is reported using PLCC correlation metrics, see Table 6.1, Table 6.2. To highlight the validity and generalizability of the

Table 6.1: Comparison of the video quality metrics for full-reference metrics, using the Pearson correlation coefficient

.

Video Dataset/PLCC for VQM	PSNR	SSIM	FovVideoVDP	PSNR-M+
CSQ	0.38	0.15	0.38	0.44
LIVE-NFLX	0.3	0.34	0.32	0.4
KoNViD-1k	-	-	-	-

Table 6.2: Comparison of the video quality metrics for non-reference metrics, using the Pearson correlation coefficient

.

Video Dataset/PLCC for VQM	VIDEVAL	BRISQUE	PaQ-2-PiQ	Proposed VQM (NRspttemVQA)
CSQ	0.26	0.21	0.19	0.48
LIVE-NFLX	0.68	0.08	0.11	0.44
KoNViD-1k	0.23	0.67	0.64	0.4

proposed NR VQA, it has been tested on three public and popular video databases: KoNViD1k [15], LIVE-Netflex [137, 138], and the proposed CSQ dataset.

LIVE Netflix QoE database consists of 112 distorted videos created from 14 undistorted videos by overlaying a set of 8 different playback templates. This Chapter used all public videos. KoNViD-1k is a subjectively annotated VQA database of 1200 public video sequences aiming for natural, authentic distortions displaying a wide range of content. This Chapter used half of the randomly selected videos. These databases are chosen because they are publicly available, have sequence lengths greater than 10 seconds, and contain video compression and transmission distortions. CSQ database is a large-scale

set of encoded videos with constant subjective assessment; the database uses H.264 compression for the 36 video sequences. Table 5.2, with an overview of existing video data sets, clearly demonstrates that the SQ database is a large-scale set. The LSVQ database is not publically available, so it could not be used [142].

The proposed metric gives the most consistent positive correlation among non-reference metrics and a comparable stable correlation with full-reference metrics. The better prediction of subjective perception is mainly due to the ability of the metric to generalize predictions for video sequences. As described above, the full-reference metric FovVideoVDP, like PSNR-M+, shows approximately the same results for video sequences with a typical bitrate. However, PSNR-M+ gives an average 15% higher correlation for video with motion. The strong artefacts presented in these datasets are evident in the low overall correlations. For example, VIDEVAL predicts a comparable correlation between prediction and real subjective ratings only for videos without significant distortion. BRISQUE, in contrast, predicts the subjective assessment of real distortions well. Which once again points to the problem of stable operation without reference metrics. The proposed video quality metric NRspttemVQA has a comparable correlation interval with full reference metrics, Fig. 4.9. Visualisation of NRspttemVQA prediction on independent databases results presented in Fig. 6.6, Fig. 6.7, Fig. 6.8, Fig 6.9. In other words, the new metric, more efficient than currently used non-reference metrics, predicts the perception of videos with different content and distortions by the human visual system.

6.3 Summary

This chapter created a new illuminance-adjusted spatio-temporal video quality score that accurately determines the local spatio-temporal video quality. It has been concluded that the proposed video quality metric achieves the high-

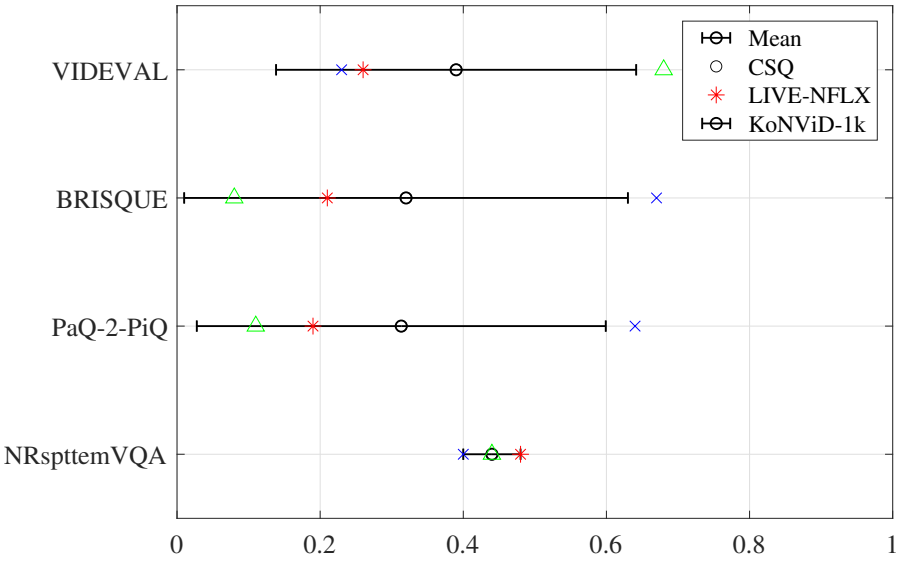


Figure 6.6: Correlation interval of non-reference video quality metrics on video sequences CSQ, LIVE-NFLX, KoNViD-1k. The new proposed metric has the most consistent high correlation of the metrics tested herein.

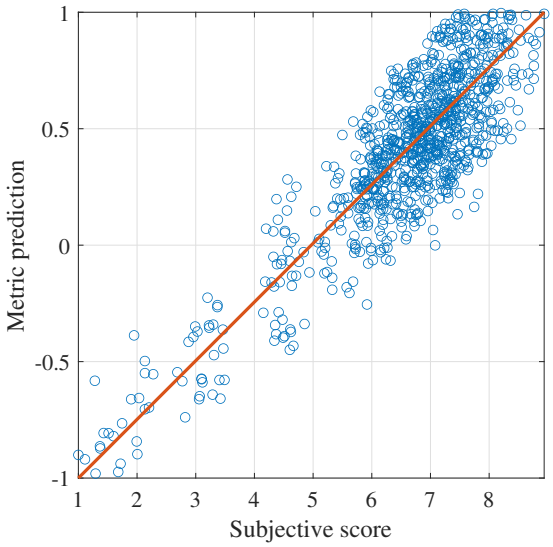


Figure 6.7: Visualisation of CSQ database results.

est prediction accuracy results for different video contents among commonly used non-reference metrics and results that are comparable to full-reference metrics. The new video quality metric can significantly advance progress on the streaming quality problem.

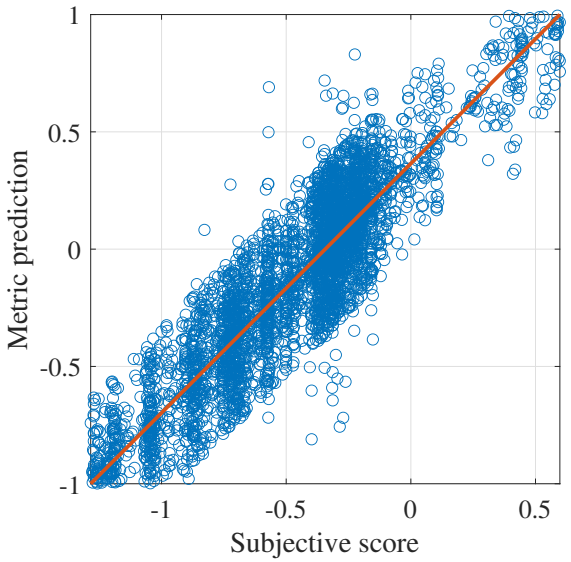


Figure 6.8: Visualisation of LIVE-NFLX database results.

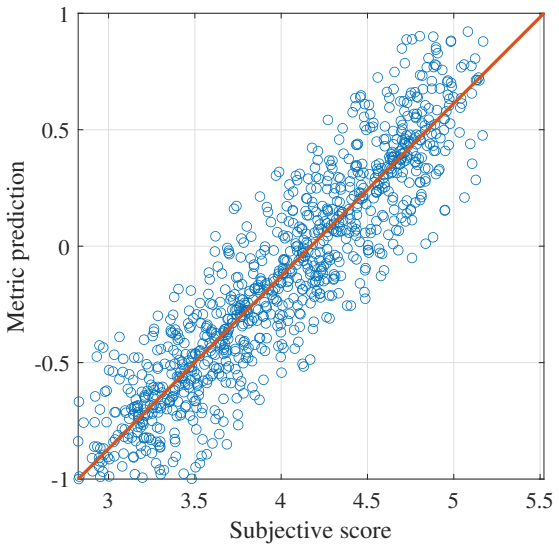


Figure 6.9: Visualisation of KoNViD-1k database results.

Chapter 7

Conclusion and Outlook

Given today's video transmission traffic, optimization of the information transmission process is a necessity. Video quality metrics used in streaming can enable new and exciting applications. An ideal video quality metric has a high correlation with quality as perceived by human observers, is standardized, and is independent of the systems or processes involved.

Over the past decade, the focus of video quality research has shifted from the broad goal of understanding how people evaluate video quality to the more limited purpose of computer design algorithms simulating human subjective assessments obtained in experiments. This work mainly focuses on new knowledge about the features of the human visual system that influence the correlation of algorithmically predicted video quality with a person's subjective perception of video quality. Using this new knowledge, the parameters of a spatiotemporal CSF model of the human visual system were refined, and a large dataset of data on the quality of distorted video showed that it is sufficient to train deep non-reference video quality metrics.

7.1 Contribution

- In Chapter 2, the architecture of modern video codecs was analyzed and problems in the encoding of video signals were identified. Chapter 3 describes the process of designing a system to measure the characteris-

tics of human visual systems that influence human perception of artefacts when viewing video content. A methodology for large-scale measurement of CSF HVS that also considers modern display devices' characteristics was created. It was created with software and hardware. Was collected and analyzed new knowledge of contrast thresholds at spatial and temporal frequencies and background pixel brightness levels. The collected visibility threshold dataset is 27840 thresholds, which is 99% larger than the largest existing dataset suitable for use in video signal processing with 209 thresholds. The resulting large-scale data is also 96% larger than the largest combined set of 11 CSF data sets.

- Chapter 4 presented a refinement of the multidimensional model of human contrast sensitivity parameters in modern video signal transmission conditions and a new full-reference video metric that expands the metric of the peak signal-to-noise ratio to features of the human visual system, allowing for a 67% improvement in stable video quality prediction.

- In Chapter 5, through collaboration with Moscow Technical University of Communications and Informatics, a new device was developed to measure the quality of users' encoded video using an acceptable minimum perceptual threshold. A new constant-quality video dataset using typical media content's dynamically changing visual quality was created. The new data collection method allows for optimisation of the time spent collecting data to create codecs and video quality metrics based on full or partial machine learning by 78%.

- Chapter 6 showed how a refined visibility model and a new constant quality video dataset were used to improve and stabilize the performance of non-reference video quality metrics. The proposed approach achieves 81% more consistent performance in predicting user subjective quality among commonly used non-reference video quality metrics and comparable consistent performance to full-reference metrics.

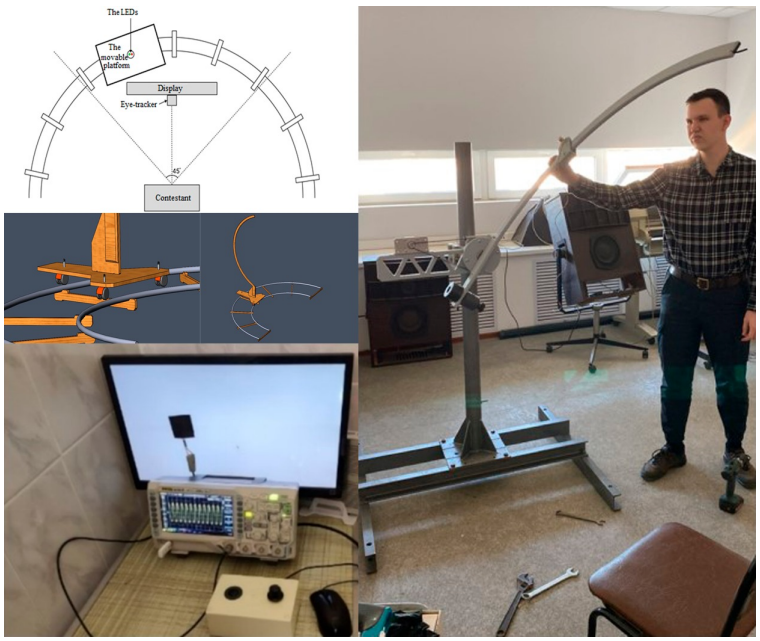


Figure 7.1: The measuring human contrast sensitivity equipment.

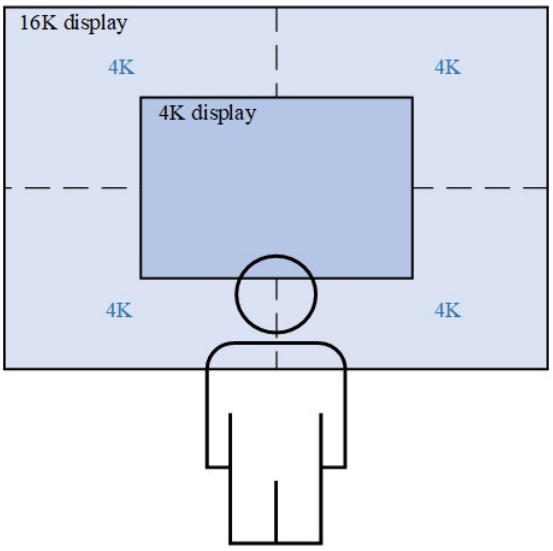


Figure 7.2: 16K resolution.

7.2 Future Work

Using the advanced signal display equipment created in this thesis, we have a unique opportunity to gain new knowledge about the functioning of the human visual system by expanding measuring methods for peripheral vision, see Fig. 7.1

This data is needed to optimize bandwidth usage for 16K video, Fig. 7.2.

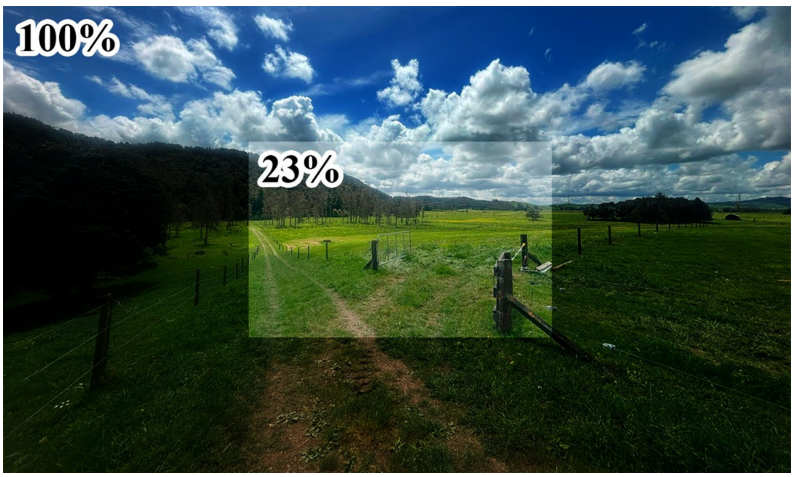


Figure 7.3: 16K technology - 100%. Modern screen -23%.

By analogy with the new knowledge obtained in this thesis about the work of the human visual system in the central area of vision, we will receive new data about the CSF in the peripheral area. Based on new knowledge and changing frame rate, a unique opportunity will open up when transmitting video to transmit only visible information, similar to the current transmission of audio signals, in which only audible frequencies are transmitted. The future work will benefit society and the economy in industries where 16K technology cannot yet be used, such as virtual reality, film and telecommunications, Fig. 7.3. In the encoding process, I'll include the contrast sensitivity function, which controls human-visible information in the video. Based on new knowledge about CSF in the peripheral region, a statistical model of visibility thresholds for spatiotemporal changes and luminance will be created. The proposed in this thesis model in the central area of HVS and future model in the peripheral region will be a filter, allowing only human-visible information.

References

[1] Cisco 2020, *Cisco Visual Networking Index 2020*, viewed 10 September 2023, https:www.cisco.com .

[2] Bampis, C, Li, Z, Katsavounidis, I, Huang, T, Ekanadham, C, Bovik, A C 2018, 'Towards Perceptually Optimized End-to-end Adaptive Video Streaming', *IEEE Transactions on Image Processing*, viewed 10 September 2023, vol. 30, pp. 5182−5197.

[3] Mohammadu, P, Ebrahimi-Moghadam, A, Shirani, A 2015, 'Subjective and Objective Quality Assessment of Image: A Survey ', *Majlesi Journal of Electrical Engineering*, vol. 9, no. 1, pp. 55−83.

[4] Mantiuk, R, Denes, G, Chapiro, A, Kaplanyan, A, Rufo, G, Bachy, R, Lian, T, Patney, A 2021, 'Fovvideovdp: a visible difference predictor for wide field-of-view video', *ACM Transactions on Graph* , vol. 40, pp. 1−19.

[5] Poynton, K 2012, *Digital Video and HD. Algorithms and Interfaces*, 2nd Edition, Morgan Kaufmann Publishers, MA, USA.

[6] Rec ITU−R BT 2341−0, *TV receiver subjective picture failure thresholds and the associated minimum quasi error free levels for good quality reception*, International Telecommunications Union, Geneva, Switzerland.

[7] Rohde, J, Schwarz, G, Co, KG 1998, *Picture Quality Measurement*, IBC, Amsterdam.

[8] Potashnikov, A, Mazin, V, Stepanov, N, Smirnov, F, Mozhaeva, A 2022, 'Analysis of Modern Methods Used to Assess the Quality of Video Sequences

During Signal Streaming', *Systems of Signals Generating and Processing in the Field of on Board Communications.*, pp. 1−4.

[9] Mozhaeva, A, Vlasuyk, I, Potashnikov, A, Streeter, L 2021, 'Full Reference Video Quality Assessment Metric on Base Human Visual System Consistent with PSNR', *28th Conference of Open Innovations Association (FRUCT)*, pp. 309−315.

[10] Hussain, A H, Lukman, A 2022, 'Forward error-correction techniques in LTE systems', *AIP Conference Proceedings*, vol. 2398, no. 1, pp. 50021

[11] Chandler, D, Phan, T, Alam, M 2015, 'Seven challenges for image quality research, in: Examining the limitations of current research into image quality assessment opens doors for further studies', *SPIE*, vol. 9014, pp. 901402.

[12] Li, Z, Aaron, A, Katsavounidis, I, Moorthy, A, Manohara, M 2004, 'Image quality assessment: from error visibility to structural similarity', *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600−612.

[13] Soundararajan R, Bovik, A C 2012, '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.

[14] Winkler, S 2012, 'Analysis of Public Image and Video Databases for Quality Assessment', *IEEE Journal of Selected Topics in Signal Processing.*, vol. 6, no. 6, pp. 616−625.

[15] Hosu, V 2017, 'The Konstanz natural video database (KoNViD-1k)', *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1−6.

[16] Mozhaeva, A, Mazin, V, Cree, M, Streeter, L 2023, 'NRspttemVQA: Real-Time Video Quality Assessment Based on the User's Visual Perception',*8th International Conference on Image and Vision Computing New*

Zealand, pp. 1−7.

[17] Mozhaeva, A, Vlasuyk, I, Potashnikov, A, Cree, M, Streeter, L 2021, 'The Method and Devices for Research the Parameters of The Human Visual System to Video Quality Assessment',*Systems of Signals Generating and Processing in the field of onboard communications*, pp. 1−5.

[18] Mozhaeva, A, Potashnikov, A, Vlasuyk, I, Streeter, L 2021, 'Constant Subjective Quality Database: The Research and Device of Generating Video Sequences of Constant Quality',*2021 International Conference on Engineering Management of Communication and Technology (EMCTECH)*, pp. 1−5.

[19] Richardson, I 2003, *H.264 and MPEG-4 Video Compression: Video Coding for Next-generation Multimedia*, England: John Wiley Sons, Chichester.

[20] Sullivan,G J, Jens, Han, O W J, Wiegand, T 2012, 'Overview of the high-efficiency video coding (HEVC) standard', *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, pp. 1649−1668.

[21] Chen, J, Liu, S, Kim, S 2019, *Algorithm description for Versatile Video Coding and Test Model 5 (VTM 5)*, 14th Meeting: Geneva, Geneva.

[22] Wieckowski, A, Ma, J, Schwarz, H, Marpe, D, Wiegand, T 2019, 'Fast partitioning decision strategies for the upcoming Versatile Video Coding (VVC) standard', *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4130−4134.

[23] Wieckowski, A, Ma, J, George, V, Schwarz, H, Marpe, D, Wiegand, T 2019, *Generalized binary splits: A versatile partitioning scheme for block-based hybrid video coding*, 2019 Picture Coding Symposium (PCS), Ningbo, China.

[24] Chen, Y, Mukherjee, D, Han, J, Grange, A, Xu, Y, Parker, S, Chen, C, Su, H, Joshi, U, Ching-Han, C, Wang, Y, Wilkins, P, Bankoski, J, Trudeau,

L, Egge, N, Valin, J, Davies, T, Midtskogen, S, Norkin, A 2020, 'An Overview of An Overview of Coding Tools in AV1: the First Video Codec from the Alliance for Open Media', *Signal and Information Processing*, vol. 9, pp. 1−15.

[25] Birman, R, Segal, Y, Hadar, O 2020, 'Multimedia Tools and Applications Overview of Research in the field of Video Compression using Deep Neural Networks', *Multimedia Tools and Applications*, vol. 79, pp. 11699−11722.

[26] Cisco 2020, *Cisco Visual Networking Index*, viewed 15 September 2023, https:www.cisco.com

[27] Jacobson, R 1995, 'An Evaluation of Image Quality Metrics', *Journal of Photographic Science*, vol. 43.

[28] Rec ITU−R BT 2023, *Methodology for the subjective assessment of the quality of television pictures*, International Telecommunications Union, Geneva, Switzerland.

[29] Rec ITU−T Rec P 910 2022, *Subjective video quality assessment methods for multimedia applications*, International Telecommunication Union, Geneva, Switzerland.

[30] Rec ITU−R Rec BT 814−4 2018, *Specification and alignment procedures for setting of brightness and contrast of displays*, Telecommunications Union, Geneva, Switzerland.

[31] Mohammadi, P, Ebrahimi-Moghadam, A, Shirani, S 2015, 'Subjective and Objective Quality Assessment of Image: A Survey', *Majlesi Journal of Electrical Engineering*, vol. 9(1), pp. 55−83.

[32] Thurstone, L 1927, 'A law of comparative judgment', *Psychological Review*, vol. 34(4), pp. 273−286.

[33] Tsukida, K, Gupta, M R 2011, *How to Analyze Paired Comparison Data*, Department of Electrical Engineering University of Washington, Seattle,

WA.

[34] Fisher, R A 1919, 'The correlation between relatives on the supposition of mendelian inheritance', *Transactions of the Royal Society of Edinburgh*, vol. 52(2), pp. 399−433.

[35] Bradley, R, Terry, M 1952, 'Rank analysis of incomplete block designs: I. the method of paired comparisons',*Biometrika*, vol.39(3/4), pp. 324−345.

[36] Wu J, Lin, W, Shi, G, Liu, A 2012, 'A perceptual quality metric with internal generative mechanism', *Proceedings of the IEEE Transactions Image Processing*, no. 99.

[37] Wang, Z, Bovik, A, Sheikh, H, Simoncelli, E 2004, 'Image quality assessment: From error visibility to structural similarity', *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600−612.

[38] Ji, G, Ni, X, Bae, H 2008, 'A Full-Reference image quality assessment algorithm based on haar wavelet transform', *IEEE Computer Society*, pp. 791−794.

[39] Ji, G, Ni, X, Bae, H 2008, 'A full-reference image quality assessment algorithm based on haar wavelet transform', *Proceedings of the International Conference on Computer Science and Software Engineering (CSSE '08)*, pp. 791−794.

[40] Cao, G, Liang, L, Ma, S, Zhao, D 2009, 'Image quality assessment using spatial frequency component', *Proceedings of the 10th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, pp. 201−211.

[41] Shi, Y, Ding, Y, Zhang, R, Li, J 2009, 'Structure and hue similarity for color image quality assessment', *Proceedings of the International Conference on Electronic Computer Technology (ICECT '09)*, pp. 329−333.

[42] Rao, D, Reddy, L 2009, 'Contrast weighted perceptual structural similarity index for image quality assessment', *Proceedings of the Annual IEEE India Council Conference (INDICON '09)*, pp. 1−4.

[43] Zhang, L, Zhang, L, Mou, X 2010, 'RFSIM: a feature based image quality assessment metric using Riesz transforms', *Proceedings of the 17th IEEE International Conference on Image Processing (ICIP '10)*, pp. 321−324.

[44] Fei, X, Xiao, L, Sun, Y, Wei, Z 2012, 'Perceptual image quality ref26-2assessment based on structural similarity and visual masking', *Signal Processing: Image Communication*, vol. 27, no. 7, pp. 772−783.

[45] Wang, Z, Bovik, A 2009, 'Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures', *IEEE Signal Processing Magazine*, vol. 26(1), pp. 98−117.

[46] Zhai, G, Zhang, W, Yang, X, Xu, Y 2005, 'Image quality assessment metrics based on multi-scale edge presentation', *Proceedings of the IEEE Workshop on Signal Processing Systems Design and Implementation (SiPS '05)*, pp. 331−336.

[47] Zhang M, Mou, X 2008, 'A psychovisual image quality metric based on multi-scale structure similarity," in Proceedings of the IEEE International Conference on Image Processing (ICIP 08)', pp. 381−384.

[48] Jin, L, Ponomarenko, N, Egiazarian, K 2011, 'Novel image quality metric based on similarity', *Proceedings of the 10th International Symposium on Signals, Circuits and Systems (ISSCS'11)*, pp. 1−4.

[49] Chou C, Hsu, Y 2011, 'Image quality assessment based on binary structure information', *Proceedings of the 7th International Conference on Computational Intelligence and Security*, pp. 1136−1140.

[50] Zhang, L, Zhang, L, Mou, X, Zhang, D 2011, 'FSIM: a feature similarity index for image quality assessment', *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378−2386.

[51] Narwaria, M, Lin, W, McLoughlin, I, Emmanuel, S, Chia, L 2012, 'Fourier transform-based scalable image quality measure', *Proceedings of the IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3364−3377.

[52] Sheikh H, Bovik, A 2006, 'Image information and visual quality', *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430−444.

[53] Shnayderman, A, Gusev, A, Eskicioglu, A 2006, 'An SVDbased grayscale image quality measure for local and global assessment', *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 422−429.

[54] Mansouri, A, Aznaveh, A, Torkamani-Azar, F, Jahanshahi, J 2009, 'Image quality assessment using the singular value decomposition theorem', *Optical Review*, vol. 16, no. 2, pp. 49−53.

[55] Narwaria M, Lin, W 2012, 'Svd-based quality metric for image and video using machine learning', *Systems, Man, and Cybernetics B*, vol. 42, no. 2, pp. 347−364.

[56] Saha, A, Bhatnagar, G, Wu, Q 2012, 'Svd filter based multiscale approach for image quality assessment', *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW '12)*, pp. 43−48.

[57] Sheikh, H, Wang, Z, Cormack, L, Bovik, A, *LIVE image quality assessment database Release 2*, viewed 10 November 2023, http://live.ece.utexas.edu/research/quality.

[58] Chandler D, Hemami, S 2007, 'VSNR: a wavelet-based visual signal-to-noise ratio for natural images', *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284−2298.

[59] Mittal, A, Moorthy A, Bovik, A 2011, 'Blind/Referenceless Image Spatial Quality Evaluator', *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pp. 723−727.

[60] Mozhaeva, A, Mazin, V, Cree, M, Streeter, L 2023, 'Video Quality Assessment Considering the Features of the Human Visual System', *mage and Vision Computing. IVCNZ 2022. Lecture Notes in Computer Science*, vol. 13836.

[61] Mozhaeva, A, Vlasuyk, I, Potashnikov, A, Streeter, L 2021, 'Full reference objective metric for assessing video quality compatible with PSNR considering the frequency and peripheral characteristics of human vision', *Digital Signal Processing Application Considerations*, vol. 2, pp. 44−54.

[62] Marziliano, P, Dufaux, F, Winkler, S, Ebrahimi, T, Sa, G 2002, 'A no-reference perceptual blur metric', *International Conference on Image Processing (ICIP '02)*, pp. 57−60.

[63] Ong, E, Lin, W, Lu Z 2003 , 'A no-reference quality metric for measuring image blur', *Proceedings of the 7th International Symposium on Signal Processing and Its Applications*, vol. 1, pp. 469−472.

[64] Dijk, J, Van Ginkel, M, Van Asselt, R, Van Vliet, L, Verbeek, P 2003, 'A new sharpness measure based on Gaussian lines and edges', *CAIP*, vol. 2756, pp. 149−156.

[65] Chung, Y, Wang, J, Bailey, R, Chen, S, Chang, S 2004, 'A non-parametric blur measure based on edge analysis for image processing applications', *Proceedings of the IEEE Conference on Cybernetics and Intelligent Systems*, pp. 356−360.

[66] Wang, Z, Bovik, A, Evans, B 2000, 'Blind measurement of blocking artifacts in images', *Proceedings of the International Conference on Image Processing (ICIP '00)*, pp. 981−984.

[67] Bovik A, Liu, S 2001, 'DCT-domain blind measurement of blocking artifacts in DCT-coded images', *Proceedings of the IEEE Interntional Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1725−1728.

[68] Meesters L, Martens, J 2002, 'A single-ended blockiness measure for JPEG-coded images', *Signal Processing*, vol. 82, no.3, pp. 369−387.

[69] Pan, F, Lin, X, Rahardja, S, Ong, E, Lin, W 2004, 'Measuring blocking artifacts using edge direction information', *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '04)*, pp. 1491−1494.

[70] Perra, C, Massidda, F, Giusto, D 2005, 'Image blockiness evaluation based on sobel operator', *Proceedings of the IEEE International Conference on Image Processing 2005*, pp. I−389.

[71] Suthaharan, S 2009, 'No-reference visually significant blocking artifact metric for natural scene images',*Signal Processing*, vol. 89, no. 8, pp. 1647−1652.

[72] Chen C, Bloom, J 2010, 'A blind reference-free blockiness measure', *Proceedings of the 11th Pacific Rim conference on Advances in Multimedia Information Processing (PCM '10)*, pp. 112−123.

[73] Ying, Z, Niu, H, Gupta, P, Mahajan, D, Ghadiyaram, D, Bovik, A 2020, 'From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality', *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3572−3582.

[74] Gao, F, Wang, Y, Li, P, Tan, M, Yu, J, Zhu, Y 2017, 'Deepsim: Deep similarity for image quality assessment',*Neurocomputing*, vol. 257, pp. 104−114.

[75] Kim, J, Lee, S 2017, 'Deep learning of human visual sensitivity in image quality assessment framework',*Computer Vision and Pattern Recognition (CVPR)*, pp. 1969−1977.

[76] Kim, J, Zeng, H, Ghadiyaram, D, Lee, S, Zhang, L, Bovik, A 2017, 'Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment',*IEEE Signal Processing*

Magazine, vol. 34(6), pp. 130−141.

[77] Li, J, Ling, S, Wang, J, Le Callet, P 2020,'A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing',*28th ACM International Conference on Multimedia*, pp. 3339−3347.

[78] Prashnani, E, Cai, J, Mostofi, Y, Sen, P 2018, 'Pieapp: Perceptual image error assessment through pairwise preference',*The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1808−1817.

[79] Zhang, R, Isola, P, Efros, A, Shechtman, E, Wang, O 2018, 'The Unreasonable Effectiveness of Deep Features as a Perceptual Metric',*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586−595.

[80] IRG-AVQA ,*Intersector Rapporteur Group Audio-visual Quality Assessment*, viewed 25 August 2023, https://www.itu.int/en/irg/avqa/Pages/default.aspx

[81] VQEG, *Visually Lossless Quality Analysis (VLQA)*, viewed 5 September 2023, ftp://vqeg.its.bldrdoc.gov/Documents/Projects/vlqa/minutes/

[82] 'History of LIVE Public-domain Subjective Picture Quality Databases', viewed 25 August 2023, https://live.ece.utexas.edu/research/Quality/history.html

[83] National Research Council 1989, *Human Performance Models for Computer Aided Engineerings*, The National Academies Press, Washington, DC.

[84] Fairchild, M 2013, *Color Appearance Models*, John Wiley & Sons, USA.

[85] Schwartz, G, Levine, J 2021, *Luminance adaptation. In Retinal Computation*, pp. 26−46.

[86] Luizov, A 1961, *Inertia of vision*, Oborongiz, Moscow.

[87] Watson, A, Ahumada, A 2016, 'The pyramid of visibility',*Journal of Vision*, vol. 16. no. 12. p. 567−567.

[88] Seshadrinathan, K, Pappas, T, Safranek, R, Chen, J, Wang, Z, Sheikh, H, Bovik, A 2009, 'Image Quality Assessment', *Ac. Press*, pp. 553−595.

[89] Cecchi, A 2018, 'Cognitive penetration of early vision in face perception',*Consciousness and Cognition*, vol. 63, pp. 254−266.

[90] Schade, O 1956, 'Optical and photoelectric analog of the eye',*Journal of Optical Society of America*, vol. 46(9), pp. 721−739.

[91] Mantiuk, R, Ashraf, M, Chapiro, A 2022, 'stelaCSF - A Unified Model of Contrast Sensitivity as the Function of Spatio-Temporal Frequency, Eccentricity, Luminance and Area',*ACM Transactions on Graphics (Proc. of SIGGRAPH 2022)*, vol. 41(4), no. 145.

[92] De Lange, H 1952, 'Experiments on flicker and some calculations on an electrical analogue of the foveal systems',*Physica*, vol. 18(11), pp. 935−950.

[93] Volkov, V, Luizov, A, Ovchinnikov, B, Travnikova, N 1989, *Ergonomics of human visual activity*, Mashinostroenie, Leningrad.

[94] Campbell, F, Cooper, G, Robson, J 1968, 'Application of Fourier analysis to the visibility of gratings',*The Journal of Physiology*, vol. 197, pp. 551−566.

[95] De Lange, H 1958, 'Research into the dynamic nature of the human fovea-cortex systems with intermittent and modulated light',*The Journal of the Optical Society of America*, vol. 48, no. 11. p. 0784−789.

[96] Daly, S 1993, 'Visible differences predictor: an algorithm for the assessment of image fidelity',*Digital Images and Human Vision*, pp. 179−206.

[97] Kelly, D 1979, 'Stabilized spatio-temporal threshold surface',*Journal of Optical Society of America*, vol. 69(10), pp. 1340−1349.

[98] Daly, S 1998, 'Engineering observations from spatiovelocity and spatiotemporal visual models',*Human Vision and Electronic Imaging*, vol. 3299, pp. 180−191.

[99] Kulikowski, J 1971, 'Some stimulus parameters affecting spatial and temporal resolution in human vision',*Vision Research*, vol. 11, pp.83−93.

[100] Van Nes, F, Bouman, M 1967, 'Spatial modulation transfer in the human eye',*The Journal of the Optical Society of America*, vol. 57, pp. 401−406.

[101] Mantiuk,R, Kim, M, Ashraf, M, Xu, Q, Luo, M, Martinovic, J, Wuerger, S 2020, 'Practical Color Contrast Sensitivity Functions for Luminance Levels up to 10000 cd/m2',*Color and Imaging Conference*, pp. 1−6.

[102] Laird, J, Rosen, M, Pelz, J, Montag, E, Daly, S 2006, 'Spatio-velocity CSF as a function of retinal velocity using unstabilized stimuli',*Human Vision and Electronic Imaging*, vol. 6057, pp. 32−43.

[103] Watson, A 2000, 'Visual detection of spatial contrast patterns: Evaluation of five simple models',*Optics Express*, vol. 6 (1), pp. 12−33.

[104] Mantiuk, R, Kim, K, Rempel, A, Heidrich, W 2011, 'HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions',*ACM Transactions on Graphics 30*, pp. 1−14.

[105] Wuerger, S, Ashraf,M, Kim, M, Martinovic, J, Pérez-Ortiz, M, Mantiuk, R 2020, 'Spatio-chromatic contrast sensitivity under mesopic and photopic light levels',*Journal of Vision*, vol. 20, pp. 23.

[106] Rovamo, J, Luntinen, O, Näsänen, R 1993, 'Modelling the dependence of contrast sensitivity on grating area and spatial frequency',*Vision Research*,

pp. 2773−2788.

[107] Robson, J 1966, 'Spatial and temporal contrast sensitivity functions of the visual system',*Journal of the Optical Society of America*, vol. 56, pp. 1141−1142.

[108] Snowden, R, Hess, R, Waugh, S 1995, 'The Processing of Temporal Modulation at Different Levels of Retinal Illuminance', *Vision Research*, vol. 35, pp. 775−789.

[109] Virsu, V, Rovamo J 1979, 'Visual resolution, contrast sensitivity, and the cortical magnification factor',*Experimental Brain Research*, vol. 37, pp. 475−494.

[110] Virsu, V, Rovamo, J, Laurinen,P 1982, 'Temporal Contrast Sensitivity and Cortical Magnification', *Vision Research*, vol. 22, pp. 1211−1217.

[111] Wright, M, Johnston, A 1983, 'Spatiotemporal contrast sensitivity and visual field locus', *Vision Research*, vol. 23, pp. 983−989.

[112] Anderson, S, Mullen, K, Hess, R 1991, 'Human peripheral spatial resolution for achromatic and chromatic stimuli: limits imposed by optical and retinal factors', *The Journal of Physiology*, vol. 442, pp. 47−64.

[113] Kontsevich, L, Tyler, C 1999, 'Bayesian adaptive estimation of psychometric slope and threshold', *Vision Research*, vol. 39(16), pp. 2729−2737.

[114] Flanagan, P, Markulev, C 2005, 'Spatio-temporal selectivity of loss of colour and luminance contrast sensitivity with multiple sclerosis and optic neuritis', *Ophthalmic Physiol Opt.*, vol. 25(1), pp. 57−65.

[115] Njeru, S, Osman, M, Brown, A 2021, 'The Effect of Test Distance on Visual Contrast Sensitivity Measured Using the Pelli-Robson Chart', *Translational Vision Science and Technology*, vol. 10 (32).

[116] García-Pérez, M, Eli, E 2001, 'Luminance artifacts of cathode-ray tube displays for vision research',*Spatial vision*, vol. 14(2), pp. 201−215.

[117] Talbot, H 1836, 'Facts relating to optical science',*The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 9(51), pp. 1−4.

[118] Kaiser, J, Schafer, R 1980, 'On the use of the I0-sinh window for spectrum analysis', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 105−107.

[119] Kudryashova, A, Adzhemov, A, Vlasuyk, I 2019, 'Application of weber-fechner law in image transmission in the field of onboard communications',*Systems of Signals Generating and Processing in the Field of on-Board Communications*, pp. 1−6.

[120] Hubel, D 1990, *Eye, brain, vision*, Mir, Moscow.

[121] Rec. ITU−R BT. 709−6 2015, *Parameter values for the HDTV standards for production and international programm exchange*, International Telecommunications Union, Geneva, Switzerland.

[122] Xu, R, Wang, H, Thibos, L, Bradley, A 2017, 'Interaction of aberrations, diffraction, and quantal fluctuations determine the impact of pupil size on visual quality',*Journal of the Optical Society of America*, vol. 34, pp 481−492.

[123] Watson, A 2013, 'Vision models and visual quality',*European Workshop on Visual Information Processing (EUVIP)*, pp. 275−279.

[124] Winn, B, Whitaker, D, Elliott, D, Phillips, N 1994, 'Factors affecting light-adapted pupil size in normal human subjects',*Invest. Ophthalmol. Vis. Sci.*, vol. 35, pp. 1132−1137.

[125] Campbell, F, Gregory, A, 'Effect of size of pupil on visual acuity',*Nature*, pp. 1121−1123.

[126] Laughlin, S 1992, 'Retinal information capacity and the function of the pupil',*Ophthalmic Physiolog.*, vol. 12, pp. 161−164.

[127] Woodhouse, J 1975, 'The effect of pupil size on grating detection at various contrast levels',*Vis. Res.*, vol. 15, pp. 645−648.

[128] Robson, J 1966, 'Spatial and temporal contrast sensitivity functions of the visual system',*Journal of the Optical Society of America*, vol. 56, pp. 1141−1142.

[129] Davis, J, Hsieh, Y, Lee, H 2015, 'Humans perceive flicker artefacts at as high as 500 H',*Scientific Reports*, vol. 5(1), pp. 1−4.

[130] Harding, G, Harding, P 2010, 'Photosensitive epilepsy and image safety',*Applied Ergonomics*, vol.41, pp. 504−508.

[131] Potashnikov, A, Vlasuyk, I, Augstkaln, I 2017, 'Analysis of methods for detecting moving objects of different types on video image',*Fundamental problems of radio electronic instrumentation*, pp. 1201−1204.

[132] Gonzalez, R, Woods, R 2012, *Digital Image Processing*, Technosphere, Moscow.

[133] Vlasuyk, I 2009, 'Development of a model of the human visual system for the method of objective image quality control in digital television systems',*Telecommunications and Transportation*, pp. 189−192.

[134] Mazin, V, Cree, M, Streeter, L, Nezhivleva, K, Mozhaeva, A 2023, 'Research and Application of the Adaptive Model of the Human Visual System for Improving the Effectiveness of Objective Video Quality Metrics',*33rd Conference of Open Innovations Association (FRUCT)*, pp. 192−197.

[135] Tsung-Jung, L, Yu-Chieh, L, Weisi L,Jay Kuo, C 2013, 'Visual quality assessment: recent developments, coding applications and future trends',*APSIPA Transactions on Signal and Information Processing*, vol.2, no.1, pp. 20.

[136] Narwaria, M, Da Silva, M, Le Callet, P 2015, 'HDR-VQM: An objective quality measure for high dynamic range video',*Signal Processing: Image Communication*, vol. 35, pp. 46−60.

[137] Bampis, C, Li, Z, Moorthy, A, Katsavounidis, I, Aaron, A, Bovik, A 2017, 'Study of Temporal Effects on Subjective Video Quality of Experience',*IEEE Transactions on Image Processing*, vol. 26, no.11, pp. 5217−5231.

[138] Bampis, C, Li, Z, Moorthy, A, Katsavounidis, I, Aaron A, Bovik, A 2016, *LIVE Netflix Video Quality of Experience Database*, viewed 10 September 2023, live.ece.utexas.edu/research/LIVE/NFLXStudy/index.html

[139] Lin, J, Song, R, Wu, C, Liu, T, Wang, H, Kuo, C 2015, 'MCL-V: A streaming video quality assessment database',*The Journal of Visual Communication and Image Representation*, vol. 30, pp. 41−49.

[140] Wang, Y, Inguva S, Adsumilli, B 2019, 'YouTube UGC dataset for video compression research',*IEEE 21st Int. Workshop Multimedia Signal Process. (MMSP)*, pp. 1−5.

[141] Mozhaeva, A, Vashenko, E, Selivanov, V, Potashnikov, A, Vlasuyk, I, Streeter, L 2022, 'Analysis Of Current Video Databases For Quality Assessment',*T-Comm*, vol. 16(2), pp.48−56.

[142] Ying, Z, Mandal, M, Ghadiyaram D, Bovik, A 2020, *LIVE Large-Scale Social Video Quality (LSVQ) Database*, viewed 7 September 2023, https://github.com/baidut/PatchVQ.

[143] VQEG 2010, *Report on the validation of video quality models for high definition video content*, viewed 11 September 2023, http://www.vqeg.org,

2010.

[144] CDVL 2010, *The Consumer Digital Video Library*,viewed 10 September 2023, http://www.cdvl.org/.

[145] Feng, X, Liu, T, Yang, D, Wang, Y 2008, *Saliency-based objective quality assessment of decoded video affected by packet losses*, Proc. Int. Conf. Image Process. (ICIP), San Diego, CA.

[146] Ramachandra, R, Goring, S, Robitza, W, Feiten, B, Raake, A 2019, 'AVT-VQDB-UHD-1: A large scale video quality database for UHD-1', *Proc. IEEE Int. Symp. Multimedia (ISM)*, pp. 1−8.

[147] Zhang, F, Moss, F, Baddeley, R, Bull, D 2018, 'BVI-HD: A Video Quality Database for HEVC Compressed and Texture Synthesised Content',*IEEE Trans. on Multimedia*, vol. 20, pp. 2620−2630.

[148] Madhusudana, P, Yu, X, Birkbeck, N, Wang, Y, Adsumilli B, Bovik, A 2020, 'Subjective and Objective Quality Assessment of High Frame Rate Videos', *arXiv*, pp. 3407−3411.

[149] Madhusudana, P, Birkbeck, N, Wang, Y, Adsumilli B, Bovik, A 2020, 'Capturing Video Frame Rate Variations through Entropic Differencing', *IEEE Signal Processing Letters*, vol. 27, pp. 1809−1813.

[150] USC Media Communications Lab 2013,*MCL-JCV Dataset*, viewed 17 September 2023, http://mcl.usc.edu/mcl- jcv-dataset/.

[151] Ying, Z, Mandal, M, Ghadiyaram D, Bovik, A 2020, 'Patch-VQ: 'Patching Up the Video Quality Problem',*CoRR*, vol. 2011.13544.

[152] Wang, H, Katsavounidis, I, Zhou, J, Park, J, Lei, S, Zhou, X, Pun, M, Jin, X, Wang, R, Wang, X, Zhang, Y, Huang, J, Kwong, S, Kuo, C 2017, 'VideoSet: A large-scale compressed video quality dataset based on JND measurement',*Journal of Visual Communication and Image Representation*,

vol.46, pp. 292−302.

[153] Seshadrinathan K, Soundararajan R, Bovik, A, Cormack, L 2010, *LIVE video quality database*,viewed 9 September 2023, http://live.ece.utexas.edu/research/quality/live$_v$*ideo.html*.

[154] Seshadrinathan, K, Soundararajan, R, Bovik, A, Cormack, L 2010, 'Study of subjective and objective quality assessment of video',*IEEE Trans. Image Process*, vol. 19, no. 6, pp. 1427−1441.

[155] Seshadrinathan, K, Soundararajan, R, Bovik A, Cormack, L 2010, 'Study of subjective and objective quality assessment of video',*IEEE Trans. Image Process.*, vol.19 (6), pp. 1427−1441.

[156] Lee, D, Paul, S, Bampis, C, Ko, H, Kim, J, Jeong, S, Homan , B, Bovik, A 2022, 'A Subjective and Objective Study of Space-Time Subsampled Video Quality', *IEEE Trans Image Process*, vol. 31, pp. 934−948.

[157] Yu, X, Birkbeck, N, Wang, Y, Bampis, C, Adsumilliand B, Bovik, A 2021, 'Predicting the Quality of CompressedVideos With Pre-Existing Distortions', *IEEE Transactions on Image Processing*, vol. 30, pp. 7511−7526.

[158] De Simone, F 2009, *EPFL-PoliMI video quality assessment database* ,viewed 9 September 2023, http://vqa.como.polimi.it/

[159] De Simone, F 2009, *Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel*, Proc. Int. Workshop Quality of Multimedia Experience (QoMEX), San Diego, CA.

[160] Pechard, S, Pepion, R, Le Callet, P 2008, *IRCCyN IVC 1080i database*, viewed 9 September 2023, http://www.irccyn.ec-nantes.fr/spip.php?article541

[161] Pechard, S, Pepion, R, Le Callet, P 2008,*Suitable methodology in subjective video quality assessment: A resolution dependent paradigm*, Proc. Int.

Workshop Image Media Quality and its Application. (IMQA), Kyoto, Japan.

[162] Boulos, F, Chen, W, Parrein, B, Le Callet, P 2009, *IRCCyN IVC SD RoI database*, viewed 10 September 2023, http://www.ir- ccyn.ec-nantes.fr/spip.php?article551

[163] Pechard, S, Pepion, R, Le Callet,P 2009, *Region-of-interest intra prediction for H.264/AVC error resilience*, Proc. Int. Conf. Image Process. (ICIP), Cairo.

[164] Zhang, F, Li, S, Ma, L, Wong, Y, Ngan, K 2011, *IVP subjective quality video database*, viewed 12 September 2023, http://ivp.ee.cuhk. edu.hk/research/database/subjective/

[165] Goldmann, L, Ebrahimi, T 2010, *3D video quality assessment*, viewed 10 September 2023, http://mmspl.epfl.ch/page38842.html

[166] Goldmann, L, De Simone, F, Ebrahimi, T 2010, *A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video*, Proc. SPIE 3D Image Process. (3DIP) and Applicat., vol. 7526, pp. 242-252.

[167] Lee, J 2010, *MMSP scalable video database*, viewed 17 September 2023, http://mmspg.epfl.ch/svd

[168] Lee, J, De Simone, F, Ebrahimi, T 2011, 'Subjective quality evaluation via paired comparison: Application to scalable video coding',*IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 882−893.

[169] Wang, Y 2008, *Poly video quality databases*, viewed 17 September 2023, http://vision.poly.edu/index.html/index.php?n=Home-Page.QualityAssessmentDatabase

[170] Ou, Y, Liu, T, Zhao, Z, Ma, Z, Wang, Y 2008, *Modeling the impact of frame rate on perceptual quality of video*, Proc. Int. Conf. Image Process. (ICIP), San Diego, CA.

[171] Ou, Y, Ma, Z, Wang, Y 2009, *A novel quality metric for compressed video considering both frame rate and quantization artifacts*, Proc. Int. Workshop Video Process. Quality Metrics (VPQM), Scottsdale, AZ.

[172] Ou, Y, Zhou, Y, Wang, Y 2010, *Perceptual quality of video with frame rate variation: A subjective study*, Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Dallas, TX.

[173] Video Quality Experts Group (VQEG) 2010, *VQEG FR-TV Phase I database*, viewed 20 September 2023, ftp://ftp.crc.ca/crc/vqeg/TestSequences/

[174] Video Quality Experts Group (VQEG) 2000, *Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment*, viewed 17 September 2023, http://www.vqeg.org/

[175] CDVL 2010, *The Consumer Digital Video Library*, viewed 15 September 2023, http://www.cdvl.org/

[176] Mackin, A, Zhang, F, Bull, D 2015,' A study of subjective video quality at various frame rates',*IEEE International Conference on Image Processing (ICIP 2015)*, pp. 3407−3411.

[177] Moorthy, A, Choi, L, Bovik A, deVeciana, G 2012, 'Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies',*IEEE Journal of Selected Topics in Signal Processing*, vol. 6, pp. 652−671.

[178] Moorthy, A, Choi, L, deVeciana, G, Bovik, A 2012, 'Mobile Video Quality Assessment Database',*IEEE ICC Workshop on Realizing Advanced Video Optimized Wireless Networks*, pp. 7055−7059.

[179] Moorthy, A, Choi, L, deVeciana, G, Bovik, A 2012, 'Subjective Analysis of Video Quality on Mobile Devices',*Sixth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, pp. 1−6.

[180] Sinno Z, Bovik, A 2019, 'Large-Scale Study of Perceptual Video Quality',*IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612−627.

[181] Sinno Z, Bovik, A 2018, 'Large Scale Subjective Video Quality Study',*2018 IEEE International Conference on Image Processing*, pp. 276−280.

[182] Sinno Z, Bovik, A 2018, *LIVE Video Quality Challenge Database*, viewed 25 September 2023, http://live.ece.utexas.edu/research/LIVEVQC/index.html, 2018

[183] Keimel, C, Habigt, J, Habigt, T, Rothbucher, M, Diepold, K 2010, 'Visual quality of current coding technologies at high definition IPTV bitrates', *Multimedia Signal Processing (MMSP)*, pp. 390−393.

[184] Wang, H 2016, 'MCL-JCV: A JND-based H.264/AVC video quality assessment dataset',*Proc. IEEE Int. Conf. Image Process. (ICIP)*, pp. 1509−1513.

[185] Egorova, A, Baryshev, R, Mozhaeva, A 2023, 'Methodology of Researching Perception Identity of Regions of Users' Interests While Viewing Streaming Video Containing Various Content and Compression Artefacts',*2023 Systems of Signals Generating and Processing in the Field of on Board Communications*, pp. 1−7.

[186] Kim, J, Han, D, Tai, Y-W 2012, 'Salient Region Detection via High-Dimensional Color Transform and Local Spatial Support',*IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 9−23.

[187] Girones, X, Julia, C, Puig, D 2013, 'Full Quadrant Approximations for the Arctangent Function',*IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 130−135.

[188] Mittal, A, Moorthy, A, Bovik, A 2012, 'No-reference image quality assessment in the spatial domain',*IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695−4708.

[189] Tu, Z, Wang, Y, Birkbeck, N, Adsumilli, B, Bovik, A 2021, 'UGC-VQA: Benchmarking blind video quality assessment for user generated content', *IEEE Transactions on Image Processing*, vol. 30, pp. 4449−4464.

[190] Mozhaeva, A, 'Video Quality Assessment Adapted For TV Signals Considering Modern Media Content Transmission Features,',*Radio engineering*, In press.

Appendix A

Extra Information

A.1 Ethical Approvals.

All experiments involving data collection with human subjects discussed in this dissertation underwent review by the ethics committee. Fig. A.1 and Fig. A.2 show the ethical approval from the University of Waikato ethics committee to measure when an image on a screen merges with the background based on various periods, amplitudes, and brightness levels pairwise comparison experiment reported in Chapter 3, HECS-20-64, and HECS-20-58. Ethical approval for the experiment reported in Chapter 5 is given in Fig. A.3, HECS-22-01.

A.2 Methodology for researching the identity of perception of regions of user interest when watching streaming videos containing various content and compression artefacts.

Most eye-tracking research focuses on identifying and analyzing users' visual attention patterns while performing specific tasks such as reading, searching, image scanning, driving, etc. The relationship between the eyes and the mind

The University of Waikato
Private Bag 3105
Hamilton, New Zealand, 3240
0800 WAIKATO (924 528)

HECS Human Ethics Committee
Brett Langley
Telephone +64 77 838 4060
Hecs-ethics@waikato.ac.nz

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

30 November 2020

Anastasia Mozhaeva

Supervisor: Lee Streeter

Re: HECS Ethics Approval of Application HREC(HECS)2020#58 "Error Analysis, Understanding and Propagation in Video Motion Coding"

Dear Anastasia:

Thank you for submitting your amended application HREC(HECS)2020#58 for ethical approval.

We are pleased to provide formal approval for your project, including the following activities:

- Run a survey using visually distorted images in a controlled environment.
- Surveys can run between 10 and 25 minutes in length
- Survey 25 people
- Identities of participants will be kept confidential and not be related connected to their results
- Data collected from the survey will be stored on university computers with a backup at MTUCI.
- Data can be used for conference, journal, and thesis publications

Please contact the committee by email (hecs-ethics@waikato.ac.nz) if you wish to make changes to your project as it unfolds, quoting your application number with your future correspondence. Any minor changes or additions to the approved research activities can be handled outside the monthly application cycle.

We wish you all the best with your research.

Kind regards,

Brett Langley, PhD
Chairperson
HECS Human Ethics Committee
University of Waikato

Figure A.1: Ethical approval HECS-20-58.

allows eye movement tracks to detect the user's areas of interest on the screen. The work in this Appendix was presented in [185] during this doctoral research project.

The University of Waikato
Private Bag 3105
Hamilton, New Zealand, 3240
0800 WAIKATO (924 528)

HECS Human Ethics Committee
Brett Langley
Telephone +64 77 838 4060
Hecs-ethics@waikato.ac.nz

17 December 2020

Anastasia Mozhaeva
Supervisor: Lee Streeter

Re: HECS Ethics Approval of Application HREC(HECS)2020#64 "Error Analysis, Understanding and Propagation in Video Motion Coding, Part 2 and Part 3"

Dear Anastasia:

Thank you for submitting your amended application HREC(HECS)2020#64 for ethical approval.

We are pleased to provide formal approval for your project, including the following activities:

Study Part 2:

- Run a study (of 40 participants) to measure pupil dilation based on room brightness within the range of standard office settings.
- Run a study to measure when an image on a screen merges with the background based on various periods, amplitudes, and brightness levels.
- The combined study should take not more than 30 minutes to complete
- Results can be used for research dissemination at conferences, journals, research thesis, and teaching resources.

Study Part 3:

- Run a study (~20 participants) showing short video clips and images on a screen
- Tests should take approximately five minutes
- Results can be used for research dissemination at conferences, journals, research thesis, and teaching resources.

Please contact the committee by email (hecs-ethics@waikato.ac.nz) if you wish to make changes to your project as it unfolds, quoting your application number with your future correspondence. Any minor changes or additions to the approved research activities can be handled outside the monthly application cycle.

We wish you all the best with your research.

Kind regards,

Brett Langley, PhD
Chairperson
HECS Human Ethics Committee
University of Waikato

Figure A.2: Ethical approval HECS-20-64.

A.2.1 A methodology.

A methodology was developed with a hardware and software complex to study the identity of perception of user areas of interest when watching streaming videos containing various content and compression artefacts. The hardware and software complex includes a computer, an eye tracker and developed software. The software records the participant's eye position coordinates and then

The University of Waikato
Private Bag 3105
Hamilton, New Zealand, 3240
0800 WAIKATO (924 528)

HECS Human Ethics Committee
Brett Langley
Telephone +64 77 838 4060
Hecs-ethics@waikato.ac.nz

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

8 February 2022

Anastasia Mozhaeva
Lee Streeter

Re: HECS Ethics Approval of Application HREC(HECS)2022#01 "Research of the constant quality video sequences" («Исследование видеопоследовательностей постоянного качества»)

Dear Anastasia:

Thank you for submitting your amended application HREC(HECS)2022#01 for ethical approval.

We are pleased to provide formal approval for your project, including the following activities:

- Recruit up to 30 participants for research of video sequences of constant quality.
- Conduct four test sessions taking place on different days in which the participant will view a video sequence consisting of 20 second clips. The participant will select an acceptable minimum video quality using the equipment. The maximum duration of one test session will be 10 minutes.
- Participants may withdraw from the study at any time during and up to one after data collection without penalty or repercussion.

Please contact the committee by email (hecs-ethics@waikato.ac.nz) if you wish to make changes to your project as it unfolds, quoting your application number with your future correspondence. Any minor changes or additions to the approved research activities can be handled outside the monthly application cycle.

We wish you all the best with your research.

Kind regards,

Brett Langley, PhD
Chairperson
HECS Human Ethics Committee
University of Waikato

Figure A.3: Ethical approval HECS-22-01.

creates gaze maps [185].

The experiment uses the Netflix video database, which is open for non-commercial use. Videos of selected source sequences are presented in YUV422 progressive format with a fixed resolution 1920×1080. This format is the most common in the world at the present stage of technology development. The frame rate is 25 frames, used in the European Union, and the duration of each

scene is more than 10 seconds, as recommended by VT-500-11. The display is a flat screen, the monitor's diagonal in the experiments is 23.6 inches, the aspect ratio is 16:9, the display is correctly configured, and the brightness and colour are calibrated using a professional exposure—meter (Spectra Cine Professional IV-A). The display calibration follows the parameters specified in the relevant test guidelines (ISO/IEC 17025). Illumination 50 lux, white colour temperature D65, stimulus diameter 0.2 m, eye tracker installed on the monitor by the operating instructions for the eye tracker model. The minimum permissible distance from the monitor to the experiment participant for the stimulus to be in the zone of clear vision is 0.872 m, and the maximum is 1.149 m. Ambient lighting was natural, as close as possible to the average in which users view media content. Viewing sessions begin with a calibration phase, allowing participants to understand how the eye tracker works immediately. The duration of a test session without a break does not exceed three minutes. To complete the test, the participant must review all sequences. At the end of each session, data is automatically collected anonymously and recorded in a spreadsheet to calculate averages [185].

A.2.2 Experiment and Results.

Two groups participated in the test experiment. In the first part of the experiment, the participant was asked a question about the video content before viewing the video, obliging the participant to look for the answer on the monitor, or in other words, involving cognitive reactions. The questions aim to program areas of interest, or in other words, the questions encourage the user to look for answers in the video. For example, the video shows a worker and the question: "What colour are work gloves?" After viewing, a window appeared where there were places for an answer; after the answer, the next video series with questions was automatically played. In the second part of the experiment, another group of participants, the participant watched the video without asking questions.

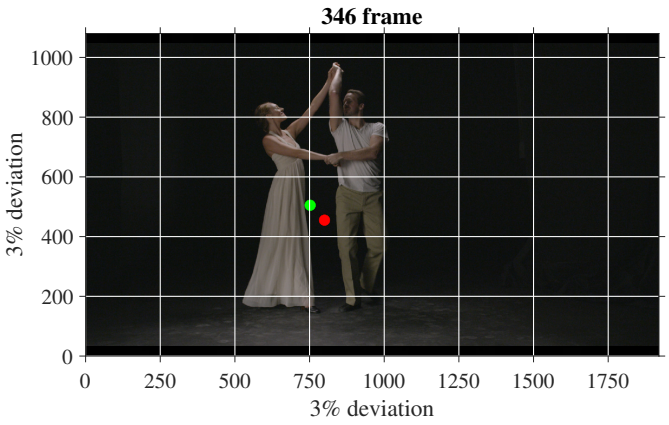


Figure A.4: Frame 346 is a fragment of a ballroom dance, where the red dot denotes the average view of the participant without a cognitive component and the green dot with a cognitive component.

A grey screen separated the video sequences. Splitting the video allows for studying changes in the perceptual identity of user interest regions of each scene. Each participant watches the same video sequence five times. The software determines eye position and focus. Technology makes it possible to study fine eye movements and visual behaviour [185]. The study was approved by the HECS Human Ethics Committee University of Waikato (HECS-20-64).

Table A.1. shows the most noticeable maximum deviations between different people of areas of interest depending on the viewed content. The repeatability of the user's areas of interest during multiple viewing of one video sequence always has deviations of no more than 5% between different people from the region of interest during the initial viewing [185].

It was found that when viewing with a cognitive component, after answering a question, participants fixed their gaze on the same area of the experiment as participants without questions, or in other words, without a cognitive component. It has been established that when multiple-viewing video sequences with different content and artefacts, the areas of interest of most users, when the task of searching for information in the frame is not set, coincide with a permissible deviation, Fig. A.4, [185].

A.3 Display model.

Increasingly, video streaming is available to viewers on various screen types. It is believed that the geometric dimensions of the screen for displaying video sequences influence the formation of a subjective assessment of the quality of perception of multimedia content on various devices. The impact of physical screen size with the effects of signal artefacts on perceived visual quality and relative depth of field has not been clear until now. The work in this Appendix was presented in [190] during this doctoral research project.

A.3.1 Stimulus and apparatus.

Dynamic changes in the visual quality of typical media content were experimentally measured depending on the physical size of the television display. The experiments were carried out according to the methodology on the hardware setup presented in Chapter 5, [18].

The proposed device consists of a screen, a projector, an external source of stimulus transmission, a manipulator for finding the minimum acceptable video quality threshold for one video [18], and a lens with a polarizing filter. The participant and projector are on the platform and can be moved with the platform to different distances. The scheme of the installation's structure is shown in Fig. A.5.

The Christie DHD800 projector projects information from an external source onto a large screen. To fix the screen parameters in each experiment, the projector provides a fixed clarity, colour reproduction, and background load.

The platform distance from the screen is x. The point on the projected image on the screen is described by coordinates (u, v) with the origin at the intersection of the optical axis with the screen. Projected image width is then $2w$ (i.e. from $u = -w$ to $u = w$) and height is $2h$ (i.e. from $v = -h$ to $v = h$). β be the half field of view (i.e. for the field of view of 30^0 used in work $\beta = 15^0$

Table A.1: The maximum deviations of results between different people by frames for all video sequences used in the test.

Title and description of the video	Database	The deviation for the first group of participants	The deviation for the second group of participants
Frame 74. Fragment from the film. (The two main characters are taken close-up)	LIVE-NFLX-II; LIVE-NFLX-I subjective video QoE database.	2%	4%
Frame 173. Fragment from a popular game called GTA		3%	6%
Frame 125. Fragment from a science fiction film, with frequent frame changes		3%	5%
Frame 857. Fragment of a flying plane. Very fast frame change		2%	3%
Frame 523. Fragment from the cooking show.		2%	4%
Frame 346. Fragment of the of a ballroom dance		5%	7%

). Then:

$$w = x \tan \beta, \tag{A.1}$$

$$h = \frac{9}{16} w = \frac{9}{16} x \tan \beta, \tag{A.2}$$

$$A = (2h)(2w) = \frac{9}{16} x^2 \tan^2 \beta, \tag{A.3}$$

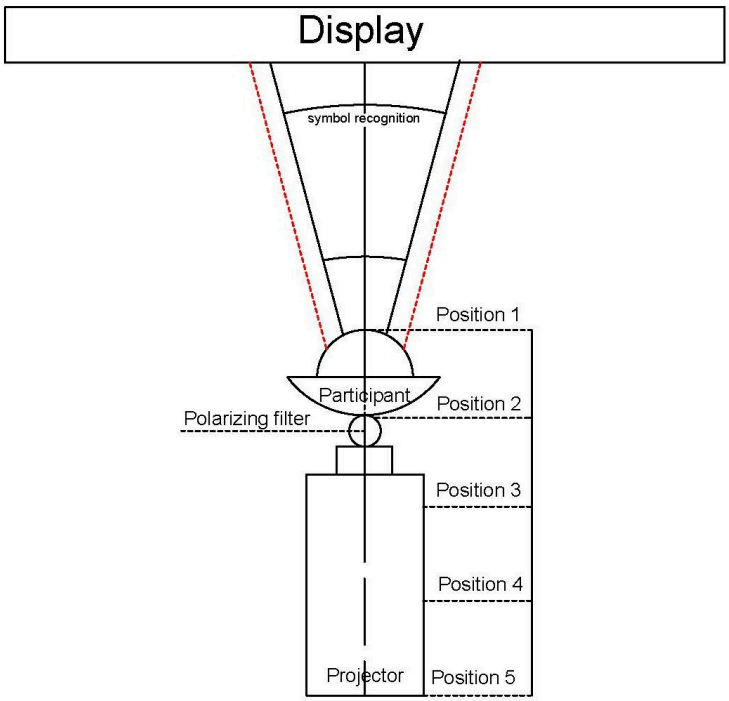


Figure A.5: Scheme of the structure of the equipment.

where A is the area of the projected image. A_s is the area of the lux meter sensor. It is fixed and independent of x. The lux meter measures intensity I_s, which is held constant. The total light emitted by the screen is, therefore,

$$I = \frac{A}{A_s} I_s = \frac{I_s}{A_s} \frac{9}{16} x^2 \tan^2 \beta, \tag{A.4}$$

The light leaving the screen towards the participant is diffusely emitted over a half hemisphere. Let the participant's pupil be radius a, thus area πa^2. The fraction of light Ω emitted by a patch at location (u, v) on the screen and received by the pupil into the eye is the area of the pupil divided by the area of the half-hemisphere of the radius of the eye from the patch. The participant's eye is at the distance $\sqrt{x^2 + u^2 + v^2}$ from a patch of screen at location (u, v). Hence:

$$\Omega = \frac{\pi a^2}{2\pi(x^2 + u^2 + v^2)}, \tag{A.5}$$

The light from a differential patch $dA = dudv$ located at (u, v) on the screen arrives obliquely to the surface of the pupil at an angle ϕ to the optical axis.

The light received at the pupil is, therefore, attenuated by

$$\cos\phi = \frac{x}{\sqrt{x^2+u^2+v^2}}, \tag{A.6}$$

The intensity of light collected by the pupil of the eye (and therefore projected onto the retina of the participant) is given by:

$$I_e = \frac{I_s}{A_s}\iint_A \Omega\cos\phi\, dA \tag{A.7}$$

Then:

$$I_e = \frac{I_s}{A_s}\iint_A \frac{\pi a^2}{2\pi(x^2+u^2+v^2)}\frac{x}{\sqrt{x^2+u^2+v^2}}\, dA, \tag{A.8}$$

$$I_e = \frac{I_s}{A_s}\iint_A f(x,u,v)\, dA \tag{A.9}$$

if

$$f(x,u,v) = \frac{\pi a^2}{2\pi(x^2+u^2+v^2)}\frac{x}{\sqrt{x^2+u^2+v^2}} \tag{A.10}$$

When the projector moves, the image changes not only in size but also in brightness. A variable density neutral density filter was made using two polarizing filters to normalise the screen's brightness. The filter's optical density variation occurs by changing the angle between the structures of the crystal lattices. According to Malus's law, the intensity of the flow after passing through 2 filters of the polarization plane, which are rotated at an angle, is:

$$I = I_0\cos^2\theta \tag{A.11}$$

As polarizing filters, a special film was used, an iodine crystal pickled between two thin layers of polyvinyl alcohol. When the projector's position was changed, the screen's surface was measured with a lux meter Testo-540 and the filter was adjusted to the illumination value adopted in the experiment, 162 lux. No matter the distance of the projector/viewing platform from the screen, the intensity of the projector is adjusted so that the intensity of light measured at the screen is being held constant. As the projector is moved backwards, the image formed on the screen increases in size as the field of view

remains constant. Therefore, from the above, the distance from the screen to the participant increases by k times when changing positions. Also, the point of the projected image on the screen is described by the coordinates (u, v) increases by k times:

$$I_2e = \frac{I_s}{A_s} \iint_{A_2} \frac{\pi a^2}{2\pi k^2(x^2+u^2+v^2)} \frac{kx}{k\sqrt{x^2+u^2+v^2}}\, dA_2, \tag{A.12}$$

where A_2 is the area of the projected image, I_2e the total intensity of light collected by the pupil of the eye in position 2:

$$A_2 = (2kh)(2kw) = k^2A, \tag{A.13}$$

Then:

$$I_2e = \frac{I_s}{A_s} \iint_{A} f(x, u, v)\, dA, \tag{A.14}$$

Therefore, the light reaching the retina is not inversely proportional to the distance of the platform from the screen. A side camera was used to measure viewing distance, and data from this camera was combined with data from a camera mounted behind the participant. The distance was measured by overlaying a grid on each keyframe so that it coincided with the centre of the screen being viewed, and then counting the number of grid lines between the screen and the viewer's eyes. For testing, we used the database Netflix [137, 138]. Each video sequence had ten different bitrates. All video sequences were separated by a grey background lasting 2 seconds.

A.3.2 Procedure and Results.

Participants in the experiment were shown the processed video sequence in several copies with different bit rates from five different distances (203, 233, 263, 293, 323 cm), simulating popular physical sizes of home television screens. Optimal viewing conditions are obtained not from the viewing angle but from the angular dimensions of the pixel, where the viewing angle for a resolution of 1920 by 1080, around 32°, makes viewing the displayed stimulus on the screen

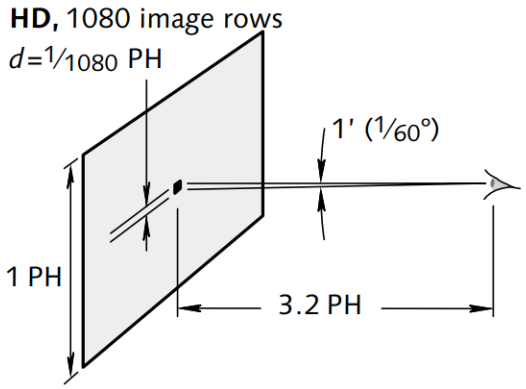


Figure A.6: View geometry [19].

optimal and most convenient, Fig. A.6. The study was approved by the HECS Human Ethics Committee University of Waikato (HECS-20-64).

Initially, video playback in the experiment for participants began with the worst bitrate. Participants had a manipulator to change the quality of perception of video content. When projecting the video, participants used a manipulator to set the threshold for the absence of distortions/artefacts of the video on the screen at their discretion. The normal environment for consuming media content was simulated as accurately as possible. Minimum perceptual thresholds were measured at five different distances in twenty-five participants. In total, 12,000 threshold values were obtained during the experiment, frame by frame. A 95% confidence interval was used to process the data.

It can be seen from Fig. A.7 that users perceive artefacts on position display screens 3 and 4 more clearly. However, since the difference in average values ranges from 0.515 to 0.600, it can be concluded that the artefacts perception of users is not significant for the provision of video content from various typical TV sizes. Therefore, the physical dimensions of a TV display device do not need to be considered for the foveation function. In other words, in the Scale Transformation block (PSNR-M+ VQM, Chapter 3), the transition from ideal viewing conditions to real ones will be constant when using screens with different diagonal sizes.

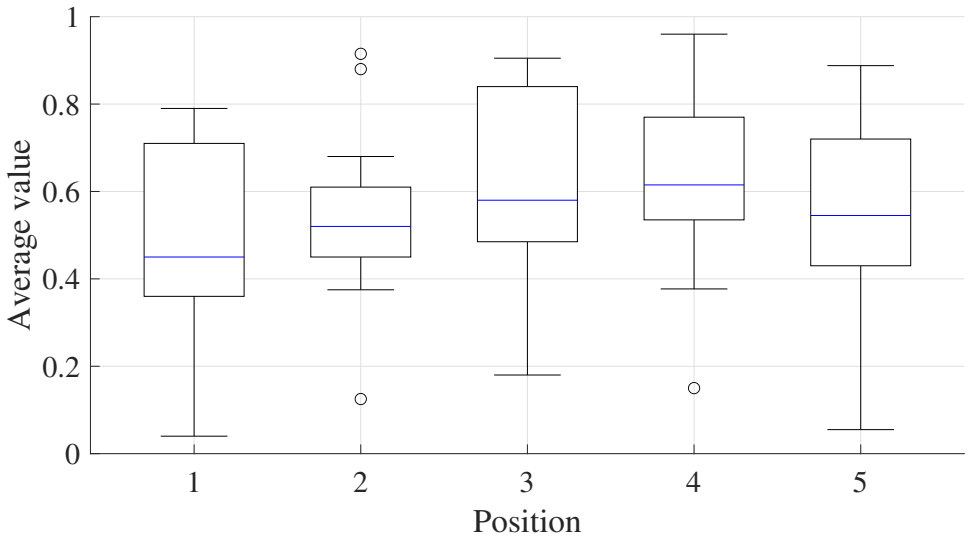


Figure A.7: The values of all positions with users' perception of quality depending on screen sizes

Appendix B

Specific Scripts

This section corresponds to Chapters 3 and 4, which proposed software for the measurement and creation model of the characteristics of human visual systems that affect the perception of artefacts by a person when viewing video content. This includes two files, one for generating the stimulus and the other for developing a polynomial approximation of the model based on the results from the subjective experiment. Also, algorithms for new video quality assessments developed in this thesis are presented.

```
import numpy as np
import pygame as pg
from scipy.special import iv


def mira_f(beta, r, y0, T):
    m = 0.5 * (iv(1, beta * np.emath.sqrt(1-r**2)) * np.sin(2*np.pi*y0/T) + 1)
    m[r >= 1] = 0.5
    m = m.T
    m = np.tile(m[:, :, np.newaxis], (1, 1, 3)) * 255
    m = m.real.astype('uint8')
    return pg.surfarray.make_surface(m)


w = 1920
h = 1080
d = 540
x, y = np.meshgrid(np.arange(0, w), np.arange(0, h))
x0 = x - w/2
y0 = y - h/2
r = 1/d * np.sqrt(x0 ** 2 + y0 ** 2)
beta = 1
sf = 2
T = 60/sf

ms = mira_f(1, r, y0, T)

pg.init()
display = pg.display.set_mode((w, h))
clock = pg.time.Clock()
q = False
while not q:
    for event in pg.event.get():
        if event.type == pg.KEYDOWN:
            if event.key == pg.K_ESCAPE:
                q = True
        if event.type == pg.MOUSEWHEEL:
            T += event.y
            ms = mira_f(1, r, y0, T)
    display.blit(ms, (0, 0))
    pg.display.update()

pg.quit()
```

Figure B.1: Generating stimulus, Python.

```
close all;
clear variables;
VIDEO_FOLDER='Z:\Study\LIVE-FBT-FCVR\';
CURRENT_VIDEO_FOLDER='2D';
REFERENCE_NAME='Coco_2d_0.yuv';
DISTORTED_NAME='Coco_2d_51.yuv';
WIDTH = 1920;
HEIGHT = 1080;

videoPlayer = vision.VideoPlayer('Position', [100 100 1920*0.6 1080*0.6]);
folder = fullfile(VIDEO_FOLDER, CURRENT_VIDEO_FOLDER);
figure


plot(continuous_subj_score)

fid = fopen(fullfile(folder, DISTORTED_NAME) ,'r');
fid_r = fopen(fullfile(folder, REFERENCE_NAME), 'r');
figure

load('H.mat')
f_max = 1 / 32;


[yr_prev, ~, ~] = yuvRead(fid_r, WIDTH, HEIGHT);
[~,~,~] = yuvRead(fid, WIDTH, HEIGHT);
PSNRm = zeros(1,Nframes);
Ws = ones(HEIGHT, WIDTH, vid_fps);
Wt = ones(HEIGHT, WIDTH, vid_fps);
Wp = ones(HEIGHT, WIDTH, vid_fps);

cur_w_idx = 1;
i = 1;
for idx = 2:Nframes
    [y,~,~] = yuvRead(fid, WIDTH, HEIGHT);
    if (BASE == 2 && is_rebuffered_bool(idx) == 1)
        continue
    end
    i = i + 1;
    [yr, ur, vr] = yuvRead(fid_r, WIDTH, HEIGHT);
    Ir = yuv2rgb(yr, ur, vr);
    yr = double(yr);
    ROI = segm(yr);
    w_s = createSpatialWeights(yr, hs);
    w_t = createTemporalWeights(yr, yr_prev,ROI, Ht, Htf, HtI, f_max, vid_fps);
    [w_p, y_roi, x_roi] = createPeriphericalWeights(yr, ROI);
    Ws(:,:,cur_w_idx) = w_s;
    Wt(:,:,cur_w_idx) = w_t;
    Wp(:,:,cur_w_idx) = w_p;
    PSNRm(i) = psnrM(double(yr), double(y), Ws, Wt, Wp, cur_w_idx);
    fprintf('Frame:%u (%u%%), PSNRm = %f, PSNR = %f, diff = %f\n', ...
        i, floor(i * 100 / Nframes), PSNRm(i), PSNR_vec(i), ...
        PSNRm(i) - PSNR_vec(i));
    yr_prev = yr;
    cur_w_idx = rem(cur_w_idx, vid_fps) + 1;
    y_roi = floor(max(y_roi,2));
    x_roi = floor(max(x_roi,2));
    Ir(y_roi-1:y_roi+1, x_roi-1:x_roi+1,:) = 0;
    Ir(y_roi, x_roi,2) = 255;
```

Figure B.2: PSNR-M+, Matlab.

```
    videoPlayer.step(Ir);
end

if (BASE == 2)
    continuous_subj_score = continuous_subj_score(~is_rebuffered_bool);
    PSNRm = PSNRm(1:N_playback_frames);
end
plot(PSNR_vec)
hold on;
plot(PSNRm);
plot(continuous_subj_score + 27);
m_corr = corr([continuous_subj_score; PSNR_vec; PSNRm;]');
```

Figure B.3: PSNR-M+, Matlab.

```
from random import random

import cv2
import numpy as np

def cost_func_mad(current_blk, ref_blk):
    return np.mean(np.abs(current_blk - ref_blk))

def motion_est_arps(img_p, img_i, mb_size, p):
    img_i = img_i.astype(float)
    img_p = img_p.astype(float)
    (row, col) = img_i.shape
    img_t = np.zeros((img_i.shape[0] + 1, img_i.shape[1] + 1))
    img_t[1:, 1:] = img_i
    img_i = img_t
    img_t = np.zeros((img_p.shape[0] + 1, img_p.shape[1] + 1))
    img_t[1:, 1:] = img_p
    img_p = img_t

    vectors = np.zeros((3, row * col // mb_size ** 2 + 1), dtype=int)
    costs = np.ones(7) * 65537
    sdsp = np.array([[0, 0, 0], [0, 0, -1], [0, -1, 0], [0, 0, 0], [0, 1, 0], [0, 0,
1]])
    ldsp = np.zeros((6, 3), dtype=int)
    check_matrix = np.zeros((2 * p + 2, 2 * p + 2))

    mb_count = 1
    for i in range(1, row - mb_size + 2, mb_size):
        for j in range(1, col - mb_size + 2, mb_size):
            x = j
            y = i

            costs[3] = cost_func_mad(img_p[i:i + mb_size, j:j + mb_size],
                                     img_i[i:i + mb_size, j:j + mb_size])
            check_matrix[p + 1, p + 1] = 1
            if j - 1 < 1:
                step_size = 2
                max_index = 5
            else:
                step_size = np.maximum(np.abs(vectors[1, mb_count - 1]),
np.abs(vectors[2, mb_count - 1]))

                if (np.abs(vectors[1, mb_count - 1]) == step_size and vectors[2,
mb_count - 1] == 0) or \
                        (np.abs(vectors[2, mb_count - 1]) == step_size and vectors[1,
mb_count - 1] == 0):
                    max_index = 5
                else:
                    max_index = 6
                    ldsp = np.append(ldsp, [[0, 0, 0]], 0)
                    ldsp[6, 1:] = np.c_[vectors[2, mb_count - 1], vectors[1, mb_count
- 1]]
            ldsp[1:6, 1:] = np.asarray([[0, -step_size], [-step_size, 0], [0, 0],
[step_size, 0], [0, step_size]])

            for k in range(1, max_index + 1):
                ref_blk_ver = y + ldsp[k, 2]
                ref_blk_hor = x + ldsp[k, 1]
                if ref_blk_ver < 1 or ref_blk_ver + mb_size - 1 > row or \
```

Figure B.4: Visibility Model, Python.

```
                    ref_blk_hor < 1 or ref_blk_hor + mb_size - 1 > col:
                continue
            if k == 3 or step_size == 0:
                continue
            costs[k] = cost_func_mad(img_p[i:i + mb_size, j:j + mb_size],
                                     img_i[ref_blk_ver:ref_blk_ver + mb_size,
                                     ref_blk_hor:ref_blk_hor + mb_size])
            check_matrix[ldsp[k, 2] + p + 1, ldsp[k, 1] + p + 1] = 1

        cost = np.min(costs[1:])
        point = np.argmin(costs[1:])
        point += 1

        # print(ldsp)
        x += ldsp[point, 1]
        y += ldsp[point, 2]
        costs = np.ones(6) * 65537
        costs[3] = cost

        done_flag = 0
        while done_flag == 0:
            for k in range(1, 6):
                ref_blk_ver = y + sdsp[k, 2]
                ref_blk_hor = x + sdsp[k, 1]
                if ref_blk_ver < 1 or ref_blk_ver + mb_size - 1 > row or \
                        ref_blk_hor < 1 or ref_blk_hor + mb_size - 1 > col:
                    continue
                if k == 3:
                    continue
                elif ref_blk_hor < j - p or ref_blk_hor > j + p or ref_blk_ver < i
- p or ref_blk_ver > i + p:
                    continue
                elif check_matrix[y - i + sdsp[k, 2] + p + 1, x - j + sdsp[k, 1] +
p + 1] == 1:
                    continue
                costs[k] = cost_func_mad(img_p[i:i + mb_size, j:j + mb_size],
                                         img_i[ref_blk_ver:ref_blk_ver + mb_size,
                                         ref_blk_hor:ref_blk_hor + mb_size])
                check_matrix[y - i + sdsp[k, 2] + p + 1, x - j + sdsp[k, 1] + p +
1] = 1

            cost = np.min(costs[1:])
            point = np.argmin(costs[1:])
            point += 1
            if point == 3:
                done_flag = 1
            else:
                x += sdsp[point, 1]
                y += sdsp[point, 2]
                costs = np.ones(6) * 65537
                costs[3] = cost

        vectors[1, mb_count] = y - i
        vectors[2, mb_count] = x - j
        mb_count += 1
        costs = np.ones(7) * 65537
        check_matrix = np.zeros((2 * p + 2, 2 * p + 2))
    return vectors[1:, 1:]
```

Figure B.5: Visibility Model, Python.

```
motion_est_arps(np.random.rand(1920, 1080), np.random.rand(1920, 1080), 3, 32)


def extract_local_features(y: np.ndarray, y_prev: np.ndarray, roi, fps):
    mb_size = 3
    vectors = motion_est_arps(y, y_prev, mb_size, 32)
    norm_vectors = normalize_motion_vectors(vectors, roi, mb_size)
    ft = norm_vectors * 1 / 32 * fps
    return get_coef_sens(ft, y / 255)


def create_arg_matrix(vec1: np.ndarray, vec2: np.ndarray, vec3: np.ndarray):
    n = vec1.size
    vec = np.c_[np.ones(n), vec1, vec2, vec3,
                vec1 ** 2, vec2 ** 2, vec3 ** 2, vec1 * vec2, vec1 * vec3, vec2 *
vec3,
                vec1 ** 3, vec2 ** 3, vec3 ** 3, vec1 ** 2 * vec2, vec1 ** 2 * vec3,
                vec2 ** 2 * vec1, vec2 ** 2 * vec3, vec3 ** 2 * vec1, vec3 ** 2 *
vec2,
                vec1 * vec2 * vec3,
                vec1 ** 4, vec2 ** 4, vec2 ** 4, vec1 ** 3 * vec2, vec1 ** 3 * vec3,
                vec2 ** 3 * vec1, vec2 ** 3 * vec3, vec3 ** 3 * vec1, vec3 ** 3 *
vec2,
                vec1 ** 2 * vec2 ** 2, vec1 ** 2 * vec3 ** 2, vec2 ** 2 * vec3 ** 2,
                vec1 ** 2 * vec2 * vec3, vec2 ** 2 * vec1 * vec3, vec3 ** 2 * vec1 *
vec2]
    return vec


def get_coef_sens(ft: np.ndarray, y: np.ndarray):
    theta_log = np.array(
        [-2.31324322336932, 1.50512703597979, 0.367349315881366, 3.18655979008766, -
7.66613420152189,
         -11.7061539477573, -6.88974431767900, 0.647889454333573, -3.86056958979751,
1.43587484657904,
         8.16356963838202, 26.5144408135850, 2.52964678897541, 10.6718916990953, -
2.33748077386246,
         -3.43703340628180, -29.0299789152317, 6.50854995409721, 14.2166716651169, -
5.90013396509646,
         -2.58449565803322, -7.19565134845494, -7.19565134848995, -9.59531656500626,
3.82098289297032,
         -0.143583643762538, 17.8683923695318, -0.713899638366362, -4.20821013444190,
1.31224899876147,
         -3.52786977368182, -1.97921613118689, 2.07038014750207, 5.38591209423837, -
0.653504566269330])
    freq_spatial_norm = 15
    freq_temp_norm = 66.6
    image_norm = 200 / 255
    y_shape = y.shape
    ft = ft.flatten().T
    y = y.flatten().T
    x = create_arg_matrix(np.zeros_like(ft), ft / freq_temp_norm, y / image_norm)
    return np.exp(np.reshape(x @ theta_log, y_shape)) / 0.22012988302511


def normalize_motion_vectors(vectors: np.ndarray, roi: np.ndarray, mb_size):
    (h, w) = roi.shape
    mb_h = np.floor(h / mb_size).astype(int)
```

Figure B.6: Visibility Model, Python.

```
    mb_w = np.floor(w / mb_size).astype(int)
    vectors = vectors[:, :mb_h * mb_w]

    vectors_y = reshape_vector_component(vectors, mb_h, mb_w, 0)
    vectors_x = reshape_vector_component(vectors, mb_h, mb_w, 1)

    vectors_y = cv2.resize(vectors_y.astype(float), dsize=None, fx=mb_size,
fy=mb_size)
    vectors_x = cv2.resize(vectors_x.astype(float), dsize=None, fx=mb_size,
fy=mb_size)

    (v_h, v_w) = vectors_y.shape
    vectors_y = np.pad(vectors_y, ((0, h - v_h), (0, w - v_w)), 'edge')
    vectors_x = np.pad(vectors_x, ((0, h - v_h), (0, w - v_w)), 'edge')

    avg_vector_x = np.mean(vectors_x[roi])
    avg_vector_y = np.mean(vectors_y[roi])

    vectors_x = np.abs(vectors_x - avg_vector_x)
    vectors_y = np.abs(vectors_y - avg_vector_y)
    return np.sqrt(vectors_y ** 2 + vectors_x ** 2)


def reshape_vector_component(vectors: np.ndarray, h, w, i):
    return np.reshape(vectors[i, :], (w, h)).transpose()
```

Figure B.7: Visibility Model, Python

```
from time import time

import cv2
import numpy as np
import scipy.io
from cv2 import VideoCapture
from scipy.io import savemat

from Models import VisibilityModel
from Models import FovealModel
from PatchesExtractors.JacobgiiSaliencyPE import JacobgiiSaliencyPE
from Predictor.RegressionPredictor import RegressionPredictor
from VideoCaptureAVI import VideoCaptureAVI
from VideoCaptureYUV import VideoCaptureYUV


class PSNRm:
    def __init__(self, video, fps, n_frames, size, verbous):
        self.predictor = RegressionPredictor()
        self.patches_extractor = JacobgiiSaliencyPE()
        self.prev_frame = None
        if video.split('.')[-1] == 'yuv':
            self.video = VideoCaptureYUV(video, size, n_frames, fps)
            if verbous:
                print("Reading YUV video...")
        else:
            self.video = VideoCaptureAVI(video, size, n_frames, fps)
            if verbous:
                print("Reading AVI video...")
        self.visibility_model = VisibilityModel
        self.foveal_model = FovealModel
        self.scores = np.zeros(self.video.n_frames)
        self.verbous = verbous

    def get_scores(self, frames_range):
        _, local_patch, _ =
self.patches_extractor.extract_patches(self.video.read(frames_range[0]))
        self.prev_frame = local_patch
        frames_range = frames_range[1:]

        for i in frames_range:
            t = time()
            frame = self.video.read(i)
            roi, local_patch, global_patch =
self.patches_extractor.extract_patches(frame)
            local_roi_mask = roi.mask[(roi.center_y + 1 - roi.height):(roi.center_y +
roi.height + 1),
                                      (roi.center_x + 1 - roi.width):(roi.center_x +
roi.width + 1)]
            local_features = self.visibility_model.extract_local_features(
                self.prev_frame, local_patch, local_roi_mask, self.video.fps)
            global_features = self.foveal_model.extract_global_features(
                global_patch, np.array([roi.center_x, roi.center_y]))
            self.scores[i] = self.predictor.predict(local_features, global_features)
            self.prev_frame = local_patch
            if self.verbous:
                print("Frame %d (%d%%): elapsed time %2.2f s, score = %2.2f"
                      % (i, round(i / (frames_range.stop - frames_range.start) * 100),
time() - t, self.scores[i]))
```

Figure B.8: NRspttemVQA, Python.

```
        return self.scores

    def get_local_patches(self, path_to_save):
        roi, local_patch, _ =
self.patches_extractor.extract_patches(self.video.read(0))
        self.prev_frame = local_patch
        frames_range = range(1, self.video.n_frames)
        local_features = np.zeros((roi.height * 2, roi.width * 2, len(frames_range)))

        for i in frames_range:
            t = time()
            roi, local_patch, _ =
self.patches_extractor.extract_patches(self.video.read(i))
            local_roi_mask = roi.mask[(roi.center_y + 1 - roi.height):(roi.center_y +
roi.height + 1),
                                      (roi.center_x + 1 - roi.width):(roi.center_x +
roi.width + 1)]
            local_features[..., i - 1] = self.visibility_model.extract_local_features(
                self.prev_frame, local_patch, local_roi_mask, self.video.fps)
            if self.verbous:
                print("Frame %d (%d%%): elapsed time %2.2f s"
                      % (i, round(i / (frames_range.stop - frames_range.start) * 100),
time() - t))
        savemat(path_to_save, {"local_features": local_features})
```

Figure B.9: NRspttemVQA, Python.