Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

No-reference Video Quality Assessment Model Based on Artifact Metrics for Digital Transmission Applications

Alexandre Fieno da Silva

Tese apresentada como requisito parcial para
conclusão do Doutorado em Informática

Orientadora
Prof.ª Dr.ª Mylène Christine Queiroz de Farias

Brasília
2017

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

No-reference Video Quality Assessment Model Based on Artifact Metrics for Digital Transmission Applications

Alexandre Fieno da Silva

Tese apresentada como requisito parcial para
conclusão do Doutorado em Informática

Prof.ª Dr.ª Mylène Christine Queiroz de Farias (Orientadora)
CIC/UnB

Prof. Dr. Alexandre de Almeida Prado Pohl
UTFPR

Prof. Dr. Bruno Luiggi Macchiavello Espinoza
UnB

Prof. Dr. Francisco Assis de Oliveira Nascimento
UnB

Prof. Dr. Li Weigang
UnB

Prof. Dr. Bruno Luiggi Macchiavello Espinoza
Coordenador do Programa de Pós-graduação em Informática

Brasília, 10 de Março de 2017

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

SAL382n

Silva, Alexandre Fieno da
No-Reference Video Quality Assessment Model Based on Artifact Metrics for Digital Transmission Applications / Alexandre Fieno da Silva; orientador Mylène Christine Queiroz de Farias. -- Brasília, 2017.
120 p.

Tese (Doutorado - Doutorado em Informática) -- Universidade de Brasília, 2017.

1. Qualidade do Conteúdo Visual. 2. Métricas Objetivas. 3. Métricas sem-referência. 4. Experimentos Subjetivos. I. Farias, Mylène Christine Queiroz de , orient. II. Título.

Acknowledgements

Completing my Ph.D. has been a hard and memorable journey in my life. However, during this process I received immeasurable help and inspiration from many people. Firstly, I am very grateful to professor Mylène C. Q. Farias for giving me this opportunity to complete my Ph.D. study, and for her support and patient during these years. Also, I would like to thank all team of Digital Signal Processing Group (GPDS), more specifically, Alessandro Silva, Pedro Garcia, Helard Becerra, and Welington Akamine who shared their experiences and knowledge with me, and have made my work at UnB an enjoyable experience. And a special thanks to Dario Morais for your supporting and partnership in this research.

I would also like to thank professor Judith A. Redi, my supervisor at Delft University of Technology (TU Delft), Delft, NL, and the team of Video Quality Experiment Group with whom I have collaborated during this research with valuable discussions during the seminars. I would like to thank all people who helped me in my Experiment: Nikita, Yuri, Chiara, Junchao, Rashi, Max, Daniel Victor, Daniel Burger, Chunyan, Marcelo, Daniele, Tingting, Alexis, Andrea, Chayan, Jiakun, Claudia, Marian, Joris, and Yi. A special thanks for Jairo, Grigori, Samur, Eduardo Souza, George, and Ricardo Ferreira that helped me a lot during my staying in Delft.

A very special thanks for my wife, Carolina, for her strong support and great sacrifice, and by having taken care of our children during my absences. To my parents and sisters, Adão, Vera, Patricia, and Paula, as well as, my mother-in-law and family, Celia, Danilo, Camila, Andrea, and Andre who always believed in me and supported me during my studies. I am in great debt with Alan for his friendship, and with Junior and Bia by their friendship and hospitality. Finally, I would like to thank my friends, Rodolfo, Isabella, Claiton, Jefferson, Gustavo Alexandre, Gustavo Neves, and all those who, directly or indirectly, helped during my Ph.D. studies.

Resumo

Um dos principais fatores para a redução da qualidade do conteúdo visual, em sistemas de imagem digital, são a presença de degradações introduzidas durante as etapas de processamento de sinais. Contudo, medir a qualidade de um vídeo implica em comparar direta ou indiretamente um vídeo de teste com o seu vídeo de referência. Na maioria das aplicações, os seres humanos são o meio mais confiável de estimar a qualidade de um vídeo. Embora mais confiáveis, estes métodos consomem tempo e são difíceis de incorporar em um serviço de controle de qualidade automatizado. Como alternativa, as métricas objectivas, ou seja, algoritmos, são geralmente usadas para estimar a qualidade de um vídeo automaticamente.

Para desenvolver uma métrica objetiva é importante entender como as características perceptuais de um conjunto de artefatos estão relacionadas com suas forças físicas e com o incômodo percebido. Então, nós estudamos as características de diferentes tipos de artefatos comumente encontrados em vídeos comprimidos (ou seja, blocado, borrado e perda-de-pacotes) por meio de experimentos psicofísicos para medir independentemente a força e o incômodo desses artefatos, quando sozinhos ou combinados no vídeo. Nós analisamos os dados obtidos desses experimentos e propomos vários modelos de qualidade baseados nas combinações das forças perceptuais de artefatos individuais e suas interações.

Inspirados pelos resultados experimentos, nós propomos uma métrica sem-referência baseada em *características* extraídas dos vídeos (por exemplo, informações DCT, a média da diferença absoluta entre blocos de uma imagem, variação da intensidade entre pixels vizinhos e atenção visual). Um modelo de regressão não-linear baseado em vetores de suporte (*Support Vector Regression*) é usado para combinar todas as *características* e estimar a qualidade do vídeo. Nossa métrica teve um desempenho muito melhor que as métricas de artefatos testadas e para algumas métricas com-referência (*full-reference*).

Palavras-chave: Qualidade do conteúdo visual, métricas objetivas, métricas sem-referência, experimentos subjetivos

Abstract

The main causes for the reducing of visual quality in digital imaging systems are the unwanted presence of degradations introduced during processing and transmission steps. However, measuring the quality of a video implies in a direct or indirect comparison between test video and reference video. In most applications, psycho-physical experiments with human subjects are the most reliable means of determining the quality of a video. Although more reliable, these methods are time consuming and difficult to incorporate into an automated quality control service. As an alternative, objective metrics, i.e. algorithms, are generally used to estimate video quality quality automatically.

To develop an objective metric, it is important understand how the perceptual characteristics of a set of artifacts are related to their physical strengths and to the perceived annoyance. Then, to study the characteristics of different types of artifacts commonly found in compressed videos (i.e. blockiness, blurriness, and packet-loss) we performed six psychophysical experiments to independently measure the strength and overall annoyance of these artifact signals when presented alone or in combination. We analyzed the data from these experiments and proposed several models for the overall annoyance based on combinations of the perceptual strengths of the individual artifact signals and their interactions.

Inspired by experimental results, we proposed a no-reference video quality metric based in several *features* extracted from the videos (e.g. DCT information, cross-correlation of sub-sampled images, average absolute differences between block image pixels, intensity variation between neighbouring pixels, and visual attention). A non-linear regression model using a support vector (SVR) technique is used to combine all *features* to obtain an overall quality estimate. Our metric performed better than the tested artifact metrics and for some full-reference metrics.

Keywords: Quality of visual content, objective metrics, no-reference metrics, subjective experiments

Contents

List of Figures

List of Tables

Abbreviations

AIC Akaike Information Criterion. 24

FR Full-Reference. 10

HVS Human Visual System. 5

INLSA Iterative Nested Least Squares Algorithm. 35

ITU International Telecommunications Union. 8

M Frame Intervals. 19

MAV Mean Annoyance Value. 23

MSE Mean Square Error. 10

MSV Mean Strength Value. 23

NR No-Reference. 10

PCC Pearson correlation coefficient. 23

PDP Percentages of Deleted Packets. 19

PSNR Peak Signal-to-Noise Error. 10

QoE Quality of Experience. 5

QoS Quality of Service. 5

RMANOVA Repeated-Measure ANOVA. 24

RR Reduced Reference. 10

SCC Spearman Rank Order Correlation Coefficient. 23

SD Standard Definition. 4

SSIM Structural Similarity and Image Quality. 11

SVR Support Vector Regression. 6

UESL Upper Empirical Similarity Limit. 25

VQEG Video Quality Experts Group. 15

VQM Video Quality Metric. 11

Chapter 1

Introduction

In modern digital imaging systems, the quality of the visual content can undergo a drastic decrease due to *impairments* introduced during capture, transmission, storage and/or display, as well as by any signal processing algorithm that may be applied to the content along the way (e.g. compression) [1, 2]. Impairments are defined as visible defects (flaws) and can be decomposed into a set of perceptual features called *artifacts* [3]. The physical signals that produce the artifacts are known as *artifact signals*. Artifacts can be very complex in their physical and perceptual descriptions [4]. Being able to detect artifacts and reduce their strength can improve the quality of the visual content prior to its delivery to the user [5].

Generally, visual quality assessment methods can be divided into two categories: subjective and objective methods. Subjective methods estimate the quality of a video by performing psychophysical experiments with human subjects [3]. They are considered the most reliable methods and are frequently used to provide *ground truth* quality scores. These methods also provide insights into mechanisms of the human visual system, inspiring, not only the design of objective quality metrics, but of all kinds of multimedia applications [6]. Nevertheless, subjective methods are expensive, time-consuming and cannot be easily incorporated into an automatic quality of service control system. On the other hand, objective methods are algorithms (metrics) that aim to predict the visual quality. Objective metrics that take into account aspects of the human visual system usually have the best performance [7, 8], but are often computationally expensive and, therefore, hardly applicable in real-time applications [9].

Designing a video quality metric that can detect impairments and estimate their annoyance (as perceived by human subjects) is not an easy task [10]. In the past decade, a big effort in the scientific community has been devoted to the development of video quality metrics that correlate well with the human perception of quality [7, 11]. Although a great number of video quality metrics has been proposed in the literature, most of these

metrics estimate impairment annoyance by comparing original and impaired videos [8,12].

Alternatives include artifact based metrics [13, 14], which estimate the strength of individual artifacts and, then, combine them to obtain an overall annoyance or quality model [14, 15]. The assumption here is that, instead of trying to estimate overall annoyance, it is easier to detect individual artifacts and estimate their strength because we 'know' their appearance and the type of process that generates them. These metrics have the advantage of being simple and not necessarily requiring the reference. They can be useful for post-processing algorithms, providing information about which artifacts need to be mitigated.

Naturally, the performance of an artifact based metric depends on the performance of the individual artifact metrics. Therefore, designing efficient artifact metrics requires a good understanding of the perceptual characteristics of each artifact, as well as a knowledge of how each artifact contributes to the overall quality [14,16]. According to Moorthy and Bovik [8], little work has been done on studying and characterizing the individual artifacts [17–19].

For example, Farias *et al.* [10,20] studied the appearance, annoyance, and detectability of common digital video compression spatial artifacts by measuring the strength and overall annoyance of these artifact signals, when presented alone or in combination in interlaced Standard Definition (SD) videos (480i). Their results showed that the presence of noisiness in videos seemed to decrease the perceived strength of other artifacts, while the addition of blurriness had the opposite effect. Moore *et al.* [21] investigated the relationships among visibility, content importance, annoyance, and strength of spatial artifacts in interlaced SD videos. Their results showed that the artifacts' annoyance are closely related to their visibility, but only weakly related to the video content.

Huynh-Thu and Ghanbari [22] examined the impact of spatio-temporal artifacts in video and their mutual interactions. They verified that spatial degradations affected the perceived quality of temporal degradations (and vice-versa). Moreover, the contribution of spatial degradations to the quality is greater than the contribution of temporal degradations. Reibman *et al.* [23] showed that the temporal artifacts (e.g. packet-loss) have an important contribution to quality and can be successfully used to predict it.

Zhai *et al.* [24] studied the perceptual quality of low bit-rate videos considering multiple dimensions. Differently from the previous works, their work does not focus on specific types of artifacts, but on different settings for video codecs, such as encoder type, video content, bit rate, frame size, and frame rate. More specifically, the authors performed a series of experiments that allowed to establish which codec settings had the greatest impact on quality. Naccari *et al.* [25], on the other hand, modeled the effects of spatial and temporal error concealment, the loss of prediction residuals, and the temporal distortion

propagation due to the motion-compensation loop.

Although a lot of work has been devoted in this study field, currently there is no clear knowledge on how different artifacts combine perceptually and how their impact depends on the physical properties of the video.

1.1 Problem Statement

The quality of a video may be altered in any of the several stages of a communication system (e.g. during compression, transmission and display). The quality of a video decreases when spatial[1] or temporal[2] artifacts are introduced [26]. It is worth pointing out that artifacts may also have spatial and temporal features, like for example packet-loss artifacts, which are a result of losses during the digital transmission [27].

Measuring the quality of a video is important for the design of communication systems that meet the minimum requirements of Quality of Service (QoS) and Quality of Experience (QoE), therefore satisfying the end-user demands [7, 8]. Like mentioned earlier, quality must be estimated taking into consideration human perception (i.e. considering aspects that are considered relevant for the Human Visual System (HVS), such as color perception, contrast sensitivity etc.). For real-time applications, it is important to design quality metrics that are sufficiently fast and computationally efficient. It is worth mentioning that, to design and validate a quality metric, it is important to compare its quality estimates with the subjective scores obtained by performing psychophysical experiments in which volunteers rate (using a predefined scale) the quality of a set of videos [26].

In the literature, most HSV-based metrics have the disadvantage of being computationally complex and, many of them, require the reference at the measurement point. For these reasons, the use of these metrics in real-time applications is impractical [16]. One possible solution is to use feature extraction methods to analyze specific characteristics or attributes of the video or image (e.g. sharpness, blur, contrast, temporal fluidity, artifacts etc.) which are considered relevant to quality. A popular type of feature extraction metrics are artifact metrics, which estimate the strength of a set of artifacts considered perceptually relevant. These artifact strengths are then combined to obtain an estimate of the quality [14, 16].

Given that measuring video quality accurately and efficiently, without the use of human subjects, is highly desirable [27], over the last twenty years a number of interesting

[1]Spatial artifacts are characterized by the presence of degradations that varying within the same frame or image, such as blocking, blurring, noise, ringing etc.

[2]Temporal artifacts are degradations that vary along the temporal domain, such as jerkiness (degradation in which the movement originally smooth and continuous is perceived as a sequence of abrupt cuts), ghosts etc.

techniques to estimate the overall video quality have been proposed [7,8,14,28,29]. Some proposals use specific functions (e.g. linear) to combine the results obtained from an analysis of common video distortions (e.g. coding and transmission artifacts) [6,30]. For example, Farias *et al.* [14] proposed a set of artifact metrics that analyzed the strength of four types of artifacts and, then, combined their results to obtain the overall perceived annoyance. Their results presented a good correlation with subjective data. Wang *et al.* [31] designed a no-reference metric for JPEG compressed images that considered two type of artifacts: blockiness and blurriness. A non-linear model based on a power function was used to combine the measurements for each artifact. Given these previous results, we believe that a system composed of different artifact metrics, individually designed to assess the levels of degradation caused by specific artifacts, can be used to produce reliable quality estimates.

1.2 Proposed Approach

In this work, we have performed six psychophysical experiments, where participants were asked to detect the artifacts and rate their annoyance and the strength of artifacts. The artifacts were presented in isolation or in combination. The goal of that set of experiments was to try to understand how individual artifact perceptual strengths combine to produce the overall annoyance and to investigate the importance of each artifact while determining the overall annoyance in impaired videos.

So, we proposed a no-reference quality assessment method for estimating the quality of videos impaired with blockiness, blurriness, and packet-loss artifacts based in several *features* extracted from de videos (e.g. DCT information, cross-correlation of sub-sampled images, average absolute differences between block image pixels, intensity variation between neighbouring pixels, and visual attention). A non-linear regression model using Support Vector Regression (SVR) is used to combine all *features* to obtain an overall quality estimate. Our metric is an improvement of the techniques currently available in the literature and is designed with the requirement of not using the reference.

In summary, the main contributions of this work are:

1. A study of the visibility and annoyance of blockiness, blurriness, and packet-loss artifacts, presented in isolation or in combination;
2. An analysis of the perceptual contribution of these artifacts to the overall quality;
3. An analysis of the perceptual contribution of these artifacts to visual attention, and;
4. The design of a no-reference quality assessment method that estimates quality by considering artifact and visual attention *features*.

1.3 Organization of the Document

This work is divided in eight Chapters. Chapter 2 describes several aspects of video quality, including objective and subjective quality assessment methods and visual attention processes. Chapter 3 presents the methodology used to perform the subjective experiments, including the physical environment, the experimental methodology, and the statistical methods used in this work. Chapters 4 to 6 discuss the results of each experiment, providing a better understanding of how the artifacts combine to produce the overall video quality. Chapter 7 describes the proposed no-reference video quality assessment method. Finally, Chapter 8 summarizes the main contributions of this work and discusses possible future works.

Chapter 2

General Aspects of Video Quality

In this chapter, we discussed the main aspects of subjective and objective quality assessment methods, describing the most common techniques, advantages, and drawbacks. We also briefly described the processes that are responsible for the human visual attention and its relationship to video quality.

2.1 Subjective Quality Assessment Methods

The most accurate way to determine the quality of a video is by measuring it using subjective quality assessment, i.e. psychophysical experiments performed with human subjects. The International Telecommunications Union (ITU) provides a set of experimental methodologies designed for image and video quality assessment. Recommendations ITU-T 930 [32] and ITU-R BT.500 [3] are the most commonly used standards for multimedia and broadcasting applications. These two documents describe settings for the physical environment and equipment, quality scoring methodologies, different ways of presenting the stimuli, and statistical techniques that can be used to analyze the subjective data. Two of the most popular methodologies are the Single Stimulus (SS) and the Double Stimulus (DS) methods.

In the SS methodology, subjects watch one test sequence at a time and the evaluation tasks are performed independently for each sequence. This way, subjects perform the experiment in a similar way to how they watch TV, that is, they do not compare the displayed video sequence with a reference (i.e. original) video. There is a variation of the SS methodology, known as Single Stimulus Continuous Quality Evaluation (SSCQE), in which the evaluation of is performed continuously. SSCQE provides a chance to analyze video quality for a more diverse and complex content, what is often difficult when using sequences of 10 seconds. Another variation of the SS methodology is the Single Stimu-

Table 2.1: Quality Category Rating (QCR) and Degradation Category Rating (DCR) scales.

Quality Category Scale		Degradation Category Scale	
Category	Score	Category	Score
Excellent	5	Imperceptible	5
Good	4	Perceptible, but not annoying	4
Fair	3	Slightly annoying	3
Poor	2	Annoying	2
Bad	1	Very annoying	1

lus Absolute Category Rating Scale (ACR-HRR), in which the (unprocessed) reference sequence is included in the experimental sessions without any identification [33].

In the DS method, the test and reference sequences are presented simultaneously to the subject, who evaluates their quality by comparing them. Similar to the SS method, the evaluation can be performed for short or continuous sequences. The variations of the DS methodologies include Double Stimulus Impairment Scale (DSIS), Double Stimulus Continuous Quality Scale (DSCQS), and Simultaneous Double Stimulus for Continuous Evaluation (SDSCE).

The scales used in the experiments can have numbers (numeric scales) and/or adjectives (nominal scales). For example, in the Absolute Category Rating (ACR) scale participants rate the quality of test sequences using a scale with five adjectives with associated numerical values. Concerning the type of *perception* being measured, when a Quality Category Rating (QCR) scale is used participants rate the video quality. Given that quality is an open ended scale (something of better or worse quality can always show up), experimenters often use a Degradation Category Rating (DCR) scale that allows participants to rate the intensity of the degradation, instead of the overall quality. Table 2.1 shows the QCR and DCR scales.

2.2 Objective Quality Assessment Methods

Unfortunately, subjective quality assessment methods are expensive, time-consuming and cannot be used in real-time applications. A solution is to use objective quality assessment methods, which are basic algorithms (i.e. implemented in hardware or in software) that perform physical signal measurements to estimate the quality of the video being displayed [11, 34]. The performance of objective methods are, frequently, estimated by comparing their results with the results gathered from subjective experiments.

Depending on the amount of reference information required by the quality assessment algorithm, objective methods can be classified in three categories: Full-Reference (FR), Reduced Reference (RR), and No-Reference (NR) [7]. FR methods require the original and test (e.g. distorted) videos to estimate quality, while RR methods requires the test video and a description or a set of parameters from the original video. Finally, NR methods only require the test video.

Since FR methods require the reference, they can only be used in off-line applications or in encoder during compression process. Most of these methods quantity the error difference between reference and test videos. Two of the most famous FR methods are the Mean Square Error (MSE) and the Peak Signal-to-Noise Error (PSNR), which can be calculated using the following equations:

$$MSE = \frac{1}{N}\sum_{i=0}^{N}(Or_i - Ds_i)^2, \tag{2.1}$$

$$PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right), \tag{2.2}$$

where N is the total number of pixels in the video frame, Or_i and Ds_i are the i^{th} pixels in the original and distorted video, respectively, and 255 is the maximum pixel intensity value.

Both PSNR and MSE have been widely used because of their physical significance and simplicity, but over the years they have been widely criticized for not correlating well with the perceived quality measurement [35,36]. One reason for this is the fact that these metrics do not incorporate aspects of the human vision system in their computation. MSE and PSNR simply perform a pixel-to-pixel comparison, without considering the content or the relationship among the pixels. They also do not consider how spatial and frequency content are perceived by human observers [27]. An example as how these metrics do not correlate well with the perceived quality measurement is depicted in Figure 2.1, where PSNR values are the same for a set of images with different impairments.

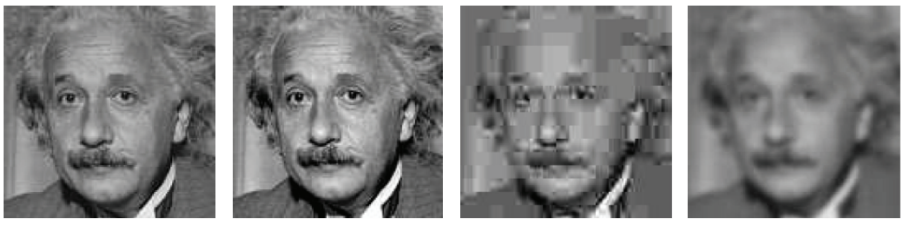


Figure 2.1: Samples of figures with different impairments and the same PSNR values: (a) original, (b) contrast stretched (26.55 dB, MSE=306), (c) JPEG compressed (26.60 dB, MSE=309), and (d) blurred (26.55 dB, MSE=308) [37].

Naturally, the FR methods that have the best performance take into account the available human vision models. Nevertheless, they are often computationally complex and require a fine temporal and spatial alignment between the reference and distorted videos. Daly *et al.* [28] proposed a metric (visible differences predictor - VDP) which estimates the visible difference between a reference and test videos, taking into account the intensity of the light, spatial frequency, and the video content. Lubin *et al.* [38, 39] proposed a multiple-scale spatial vision model for estimating the probability of detecting artifacts by analyzing their color information and temporal variations.

Wang *et al.* [29] proposed an algorithm, Structural Similarity and Image Quality (SSIM), that takes advantage of the fact that human vision is highly adapted to scene structures. In other words, the algorithm compares the local pattern normalized pixel intensity with the luminance and contrast. The metric returns values between 0 and 1, with lower values corresponding to lower quality and higher values to higher quality. Pinson *et al.* [40] proposed a Video Quality Metric (VQM) that has been adopted by the American National Standards Institute (ANSI) as a standard for objective video quality metrics. VQM measures the strength of several kind of artifacts, such as blockiness, blurriness, etc.

As mentioned earlier, RR methods require only some information about the reference video. Frequently, this information is a set of features extracted from the original and sent over an auxiliary channel. RR methods can be less accurate than FR metrics, but they are less complex. Gunawan *et al.* [41] proposed an RR method that takes into account features of local harmonic strengths (LHS). LHS is calculated taking into account the harmonic gain and loss, which are obtained from a discriminative analysis obtained from gradient images. The harmonic intensity can be introduced as a spatial activity measurement estimated from the vertical and horizontal edges of the image. Other RR metrics include works of Carnec *et al.* [42], Voran *et al.* [43], and Wolf *et al.* [44].

NR methods are designed to blindly estimate the quality of a video. Therefore, NR are more adequate for broadcasting and multimedia applications, which often require a real-time computation. Nevertheless, the design of NR is a challenge and, although several works have been proposed in the literature [45–50], a lot work still needs to be done for these methods to become reliable. One popular approach in the design of NR methods are the feature-based methods or, more specifically, the artifact-based methods [15, 51]. These methods estimate the strength of individual artifacts and combine the results to obtain an overall annoyance (or quality) score [4, 16, 31, 52].

Artifact quality assessment methods have the advantage of being simple. Also, they can provide information about which artifacts or attributes need to be enhanced in the video. However, designing artifact metrics requires a good understanding of the perceptual

characteristics of each artifact, as well as the knowledge of how each artifact contributes to the overall quality [14, 16, 45, 46]. It is worth noting that most digital videos can be affected by several artifacts that relate to each other, what makes difficult to individually perceive these artifacts and obtain an estimate of the overall quality of the video as, for example, the effects of the simultaneous presence of several artifacts. The quality of a video affected by a combination of artifacts cannot be obtained by a simple linear combination of the estimated intensities of each type of artifact. Masking and other effects of interaction between artifacts may occur, making the prediction strategy more complex and dependent on the artifacts involved [27].

In our work, we considered the most traditional artifact metrics in the literature for estimating the strength of blockiness, blurriness, and packet-loss artifacts (see Table 2.2). Wang *et al.* [31] proposed a metric that estimates blockiness by taking the difference of pixel intensities across block boundaries. Vlachos [53] proposed a method that estimates blockiness by measuring the ratio between the correlation of intra- and inter-block pixels. The algorithm split the frame into 8×8 blocks and simultaneously sampled it in the vertical and horizontal directions, assuming that all visible blockiness artifacts have a visible corner. Farias *et al.* [14] modified the metric proposed by Vlachos [53]. They considered only one of the borders of the blocking structure and, instead of down-sampling the frame simultaneously, they split it into two separate parts (i.e. into vertical and horizontal directions). They claimed this modification improves the performance of the algorithm, although it slightly increases its complexity. These algorithms are computationally efficient since they do not use complex transforms or require storing the entire image in memory.

Marziliano *et al.* [52] proposed a method that estimates blurriness by measuring the width of strong edges. Narverkar *et al.* [54] proposed a probabilistic framework that is based on the human sensitivity to blurriness in regions with different contrast levels. Crete *et al.* [55] proposed an approach that discriminates different (perceptible) levels of blurriness for the same image. In other words, their algorithm estimates quality by measuring the differences between a blurred image and its re-blurred version (higher intensity of blurriness).

Babu *et al.* [13] proposed a packet-loss metric that analyzes the differences between pixels at the macroblocks boundaries. On the other hand, Rui *et al.* [56]'s metric estimate the annoyance caused by packet-loss artifacts by measuring the attributes (length and strength) of sharp intensity discontinuities.

Table 2.2: Artifact metrics by distortions.

Artifact metrics	Artifact	Alias
Babu [13]	Packet-loss	$Pack_B$
Xia Rui [56]	Packet-loss	$Pack_R$
Marziliano [52]	Blurriness	$Blur_M$
Narverkar [54]	Blurriness	$Blur_N$
Crete [55]	Blurriness	$Blur_C$
Farias [14]	Blockiness	$Bloc_F$
Wang [46]	Blockiness	$Bloc_W$

2.3 Visual attention

Recent studies show that video quality is closely tied to gaze deployment [57]. When observing a scene, the human eye typically scans the video neglecting areas carrying little information, while focusing on visually important regions [58]. Wang *et al.* [59] showed that, within the first 2,000ms of observation, gaze patterns target the main objects in a scene. Later, the gaze is redirected to other salient areas, yet not visually important. This result suggests that visual coding should be focused, at first, into the main objects of the scene. Nevertheless, the presence of artifacts may disrupt natural gaze patterns, causing annoyance and, consequently, lower quality judgments [60]. Therefore, saliency information should be incorporated into the design of video quality metrics.

Research in the area of visual quality have focused on trying to incorporate gaze pattern information into the design of visual quality metrics [61], mostly using the assumption that visual distortions appearing in less salient areas might be less visible and, therefore, less annoying [62]. However, while some researchers report that the incorporation of gaze pattern information increases the performance of quality metrics, others report no or very little improvement [63]. One possible reason for such disagreement is that, still, the role played by visual attention in quality evaluation is unclear. Although it has been shown that, for images, artifacts in visually important regions are far more annoying than those in the background [64], it is still not clear if artifacts can create saliency (and therefore, attract gaze) on their own. And if so, it is unclear which type of artifacts can create saliency and at what perceptual strength. If artifacts can disrupt gaze patterns by creating saliency, this should be taken into account in the design of quality metrics that make use of saliency or gaze pattern information. Unfortunately, the existing knowledge in this direction is scattered.

Ninassi *et al.* [65] studied the viewing behavior during both free-viewing and quality

assessment tasks. They found that the quality task has a significant effect on the fixation duration, which increased on unimpaired images during a quality scoring task. Also, the type of impairment caused modifications in the gaze patterns. Redi *et al.* [60] analyzed the impact of three kinds of artifacts (JPEG compression, white noise and gaussian blur) and showed that they caused changes in the gaze patterns during both quality assessment and free-viewing tasks. Also, in Redi *et al.* [66] gaze pattern deviations were measured by analyzing similarities among saliency maps. They reported that the differences between the saliency maps (both for free-viewing and quality assessment tasks) seem to be more related to the strength of the artifacts impairing the images than to the type of artifacts.

With respect to video, Le Meur *et al.* [67] examined the viewing behavior during quality assessment and free-viewing tasks. Differently from images, they found that the average fixation duration is almost the same for both tasks, whereas saliency does not change significantly when videos are impaired (coding artifacts). Redi *et al.* [68] investigated to what extent the presence of packet-loss artifacts influences viewing behavior. Contrary to Le Meur *et al.* [67], they found that saliency could significantly change from free-viewing to quality assessment tasks and that these changes were related to both video content and to packet-loss annoyance. Similarly, Mantel *et al.* [69] found a positive correlation between coding artifacts annoyance and fixation dispersion.

From these results, it seems that, for both images and videos, some artifacts (e.g. packet-loss) may be able to divert gaze and viewing behavior from their natural paths. But, it is yet unclear when and how this happens. It is important to point out that most studies have focused on analyzing the impact that artifacts in isolation have on gaze patterns, like for example blockiness [69, 70] or packet-loss [68, 71]. In real-life situations, it is very likely that different artifacts are co-present in a video. For example, packet-loss may occur in the transmission of a severely compressed video, creating perceptual degradations that are very different from the single artifacts in isolation. To the best of our knowledge, there is no study that explores the impact of combinations of artifacts on gaze patterns and viewing behavior.

Chapter 3

Experimental Methodology

To understand the relationship between the perceptual strengths of blockiness, blurriness, and packet-loss artifacts and how they can be combined to estimate the overall annoyance, we performed a set of psychophysical experiments using test sequences with combinations of these artifacts at different strengths [34]. The experiments shared identical experimental methodology, interface, protocol, and viewing conditions. In this chapter, we detailed each experiment performed in this work.

3.1 Stimuli

We used seven high definition original videos chosen with the goal of generating a diverse content. The videos have a spatial resolution of 1280×720 pixels, a temporal resolution of 50 frames per second (fps), and a duration of 10 seconds. Representative frames of each original are shown in Figure 3.1. We followed the recommendations detailed in the Final Report of Video Quality Experts Group (VQEG) on the validation of objective models multimedia quality assessment (Phase I) [72], which suggest to use a set of video sequences with a good distribution of spatial and temporal properties [73]. Figure 3.2 shows the spatial and temporal activity measures of the originals.

To add artifacts to the originals, we used a system for generating artifacts [20] that allowed a control of the artifact combination, visibility, and strength, which would be impossible when using, for example, a H.264 codec[1].

Packet-loss is a distortion caused by a complete loss of the packet being transmitted, without the error concealment algorithm (at the decoder) being able to recover the missing data. These artifacts are visually characterized by the presence of rectangular areas, whose content differs from the content of the surrounding areas [68] (see Figure 3.3 (a)).

[1]As a contribution of this project, a diverse high-definition (720p) video database is made publicly available [74].

Figure 3.1: Frame videos. Top row: Park Joy. Middle row: Into Tree, Park Run, and Romeo and Juliet. Bottom row: Cactus, Basketball, and Barbecue.

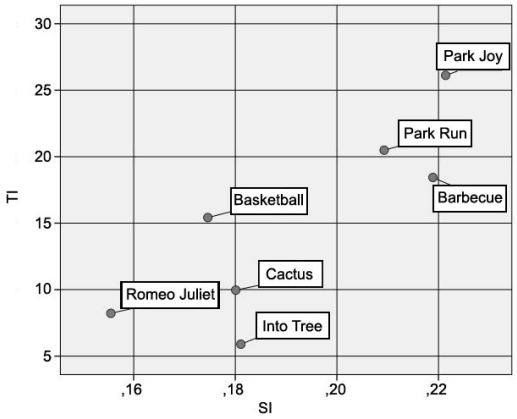


Figure 3.2: Temporal and spatial information.

To generate packet-loss artifacts, we first compressed the videos at high compression rates, what avoids inserting additional artifacts. Then, packets from the coded video bitstream were randomly deleted using different loss percentages (the higher the percentage, the lower the quality) [34]. To vary the time interval between consecutive artifacts, we changed the number of frames between I-frames.

Blockiness is a distortion characterized by the appearance of the underlying block encoding structure, often caused by a coarse quantization of the spatial frequency components during compression [32] (see Figure 3.3 (b)). To add blockiness to each video frame in our dataset, we calculated the average value of each 8×8 block of the frame and of the 24×24 surrounding block, then added the difference between these two averages to the block.

Blurriness is defined as a loss of spatial details or a reduction of edge sharpness, being more visible in textured areas or around scene objects (see Figure 3.3 (c)). To generate

(a) (b) (c)

Figure 3.3: A video frame with (a) packet-loss, (b) blockiness, and (c) blurriness artifacts.

blurriness, we used a simple low-pass filter, as suggested by Recommendation P.930 [32]. Although we can vary the filter sizes and the cut-off frequencies to control the amount of blurriness, we used a simple 5×5 moving average filter. We generated test sequences with combinations of blockiness and blurriness by linearly combining the original video with blockiness and blurriness artifact signals in different proportions (i.e. 0.4, 0.6, and 0.8) [75].

3.2 Methodology and Equipment

The experiments were performed using a PC computer with test sequences displayed on a Samsung LCD monitor of 23 inches (Sync Master XL2370HD) with resolution 1920×1080 @60hz (FullHD 1080p). The dynamic contrast of the monitor was turned off, the contrast was set at 100, and the brightness at 50. The monitor measured gamma values for luminance, red, green, and blue were 1.937, 1.566, 1.908, and 1.172, respectively. We set a constant illumination of approximately 70 lux. While watching the video sequences, the eye-movements were recorded using a SensoMotoric Instruments REDII Eye Tracker with a sampling rate of 50/60Hz. It has a pupil tracking resolution of 0.1° and a gaze position accuracy of 0.5 to 1. Participants were kept at a fixed distance of 0.7 meters from the monitor using a chinrest. The experimental methodology was the single-stimulus with hidden reference, with a 100-point continuous-scale [3, 34].

The participants were mostly graduate students from UnB and Delft University. They were considered naive of most kinds of digital video defects and the associated terminology. No vision test was performed, but participants were asked to wear glasses or contact lenses if they needed them to watch TV. The experiment started after a brief oral introduction.

The experiments were performed using the same experimental methodology. All experiments were divided into calibration, free-viewing, training, practice and an experimental session sessions:

- **Calibration:** participants were requested to focus on different points spread over the monitor screen, and their eye fixations are recorded to calibrate the eye-tracking

data;

- **Free viewing:** participants were asked to freely look at seven high quality videos, as if they are watching TV at home;
- **Training:** participants were showed all four high quality videos. Then, videos with the strongest defect derived from each of the four high quality videos are shown. The intent of this stage is to familiarize the test participants with the endpoints of the annoyance scale and to clarify the experimental task;
- **Practice:** participants ran through a limited number of practice trials. The practice trials gives the participant a chance to work through the data entry procedure and shake out last minute questions or concerns. Also, since the initial responses may be somewhat erratic, the practice stage allows the test participant responses to stabilize. No data is collected during this task;
- **Experimental Session:** Participants are asked to watch several test sequences. After each sequence is played, the participant is asked: *Did you perceive any impairments or defects in the video?*, prompting for a *Yes* or *No* answer. Then, participants are asked to perform an *annoyance* or a *strength* task. The annoyance task requires that the participant gives a numerical judgment of how annoying the detected impairment is. Impairments as annoying as those seen in the training session should be given a *100* annoyance score, sequences half as annoying a *50* annoyance score, and so on. The strength task requires that the participant rate the strength of each artifact identified in the video. Artifacts as strong as those seen in the training session should be given a *100* strength score, sequences half as strong a *50* strength score, and so on. The number of artifacts present in the test sequences varied for each experiment. To avoid fatigue, the experimental session was broken into sub-sessions, between which participants could take a break for as long as they wanted to. All experimental sessions lasted between 45 and 60 minutes.

3.3 Subjective Experiments Details

In this section, we detail the experiments performed in this work.

3.3.1 Experiment 1

In this experiment, 16 participants rated the annoyance (Exp.1a) whilst 14 participants performed strength (i.e. intensity of degradation) tasks (Exp.1s) on test sequences containing only packet-loss artifacts. To vary the strength of the artifacts, we randomly

Table 3.1: Exp. 1: Combinations of the parameters *PDP* and *M* used for each of the 7 originals.

Comb	M	PDP	Comb	M	PDP	Comb	M	PDP
1	4	0.7	5	8	0.7	9	12	0.7
2	4	2.6	6	8	2.6	10	12	2.6
3	4	4.3	7	8	4.3	11	12	4.3
4	4	8.1	8	8	8.1	12	12	8.1

deleted packets from the coded video bitstream. The *Percentages of Deleted Packets (PDP)* used were 0.7%, 2.6%, 4.3%, and 8.1%. To vary the time interval between introduced artifacts, we varied the number of frames between the I-frames. Three *Frame Intervals (M)* were used: 4, 8 and 12. The set of *PDP* and *M* parameters used in the experiments are given in Table 3.1. A total of 7 originals and 12 parameter combinations were used, resulting in $12 \times 7 + 7 = 91$ test sequences. To avoid fatigue, these videos were evaluated in a single experimental session, divided in three sub-sessions with two 10-minutes breaks.

3.3.2 Experiment 2

In this experiment, 16 participants rated annoyance (Exp.2a) whilst 15 participants performed strength tasks (Exp.2s) on test sequences containing different strengths of blockiness and blurriness artifacts, presented in isolation or in combination. Strength combinations are represented by a vector (bloc; blur), where *bloc* is the blockiness strength and *blur* is the blurriness strength. The experiments contained a set of videos with all possible combinations of the two artifact types at strengths 0.0, 0.4, and 0.6. Two additional combinations, consisting of pure blockiness and pure blurriness at strength 0.8, were added to the experiments. Table 3.2 shows all combinations used in the experiments. A total of 7 originals and 10 combinations were used in this experiments, resulting in $10 \times 7 + 7 = 77$ test sequences.

3.3.3 Experiment 3

In this experiment, 23 participants rated annoyance (Exp.3a) whilst 35 participants performed strength tasks (Exp.3s) on test sequences containing different strengths of blockiness, blurriness, and packet-loss artifacts, presented in combinations. The strength combinations are represented as a vector (PDP;bloc;blur), where *PDP* is the level of packet-loss strength, *bloc* is the level of blockiness strength, and *blur* is the level of blurriness strength.

Table 3.2: Exp. 2: Set of combinations used for each of the 7 originals: *bloc* and *blur* correspond to the blockiness and blurriness strengths, respectively.

Comb	(bloc;blur)	Comb	(bloc;blur)	Comb	(bloc;blur)
1	(0.0;0.0)	5	(0.4;0.4)	9	(0.6;0.6)
2	(0.0;0.4)	6	(0.4;0.6)	10	(0.0;0.8)
3	(0.0;0.6)	7	(0.6;0.0)	11	(0.8;0.0)
4	(0.4;0.0)	8	(0.6;0.4)		

Table 3.3: Exp. 3: Combinations for each original: *bloc* corresponds to the blockiness strength, *blur* to the blurriness strength, and *PDP* to the packet-loss ratio.

Comb.	(PDP;Bloc;Blur)	Comb.	(PDP;Bloc;Blur)	Comb.	(PDP;Bloc;Blur)
1	(0.0;0.0;0.0)	8	(8.1;0.0;0.6)	15	(0.7;0.6;0.0)
2	(0.0;0.6;0.0)	9	(0.7;0.4;0.0)	16	(8.1;0.6;0.0)
3	(0.0;0.0;0.6)	10	(8.1;0.4;0.0)	17	(0.7;0.6;0.4)
4	(8.1;0.0;0.0)	11	(0.7;0.4;0.4)	18	(8.1;0.6;0.4)
5	(0.7;0.0;0.4)	12	(8.1;0.4;0.4)	19	(0.7;0.6;0.6)
6	(8.1;0.0;0.4)	13	(0.7;0.4;0.6)	20	(8.1;0.6;0.6)
7	(0.7;0.0;0.6)	14	(8.1;0.4;0.6)		

Considering the results from the previous experiments, we selected a subset of artifact strength values to limit the number of artifact combinations. For packet-loss ratio, we chose $M = 12$ because this was the most realistic setting for the group of picture (GOP) size, which is recommended to be at most half of the frame rate. Also, we chose $PDP = 0.7\%$ and 8.1% because these values corresponded to the highest differences in annoyance (as shown in the analysis of Exp.1a). With respect to blockiness and blurriness, we chose strength values equal to 0.4 and 0.6, which were considered to be more representative of these artifacts (as shown in the analysis of Exp.2a). Table 3.3 shows all combinations used in the experiments, which include three strengths for each artifact type. Again, 7 originals and 19 combinations were used, resulting in $19 \times 7 + 7 = 140$ test sequences. To avoid fatigue, these videos were evaluated in a single experimental session, divided in three sub-sessions by two 10-minutes breaks.

3.4 Other Video Databases

Although the video database used in our experiments have a large amount of combinations of impairments, we also chose three publicly available video quality databases for testing our models: Image and Video Processing Laboratory (IVPL) database, Laboratory for Image & Video Engineering (LIVE) [76, 77] database, and Computational and Subjective Image Quality (CSIQ) Video Quality [78] database.

3.4.1 Image and Video Processing Laboratory (IVPL)

The IVPL database was developed at the Chinese University of Hong Kong. It contains 10 pristine videos and 4 types of distortions: MPEG2 compression (MPEG2), Dirac wavelet compression (Dirac), H.264 compression (H264), and packet-loss on the H.264 streaming through IP networks (IP). In this work, we have only used sequences with H264, MPEG2, and IP distortions. All videos are in raw YUV420 format, with a spatial resolution of 1920×1088 pixels, duration of 10 seconds, and a temporal resolution of 25fps. Each video was rated by 42 participants in a single-stimulus quality scale test method (ACR). Figure 3.4 shows a frame of each original video of the IVPL database.

Figure 3.4: Sample images of source video contents from IVPL database.

3.4.2 Laboratory for Image & Video Engineering (LIVE)

The LIVE Video Quality Database was developed at the University of Texas at Austin. It contains 10 pristine videos and 150 distorted videos (15 distorted videos per reference) with four different distortion types: MPEG-2 compression, H.264 compression, IP compression (i.e. simulated transmission of H.264 compressed bitstreams through error-prone IP networks), and Wireless compression (i.e. through error-prone wireless networks). Distortion strengths were adjusted to ensure that different distorted videos were separated by perceptual levels of distortion. All videos are in raw YUV420 format, with a spatial resolution of 768×432 pixels, a duration of 10 seconds, and a temporal resolution from 25 to 50 fps. Each video was assessed by 38 participants in a single stimulus study with

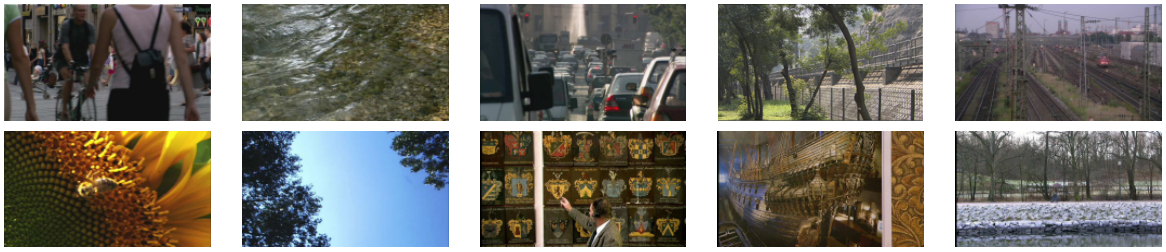


Figure 3.5: Sample images of source video contents from LIVE database.

hidden reference removal. Participants scored the video quality on a continuous quality scale. Figure 3.5 shows a frame of each video used in our tests.

3.4.3 Computational and Subjective Image Quality (CSIQ)

The CSIQ video database was developed to provide a useful dataset for the validation of objective video quality assessment algorithms. It consists of 12 high-quality reference videos and 216 distorted videos, containing six types of distortion at three different levels of distortion. The distortion types consist of four compression-based distortion types and two transmission-based distortion types, such as H.264 compression (H264), HEVC/H.265 compression (HEVC), Motion JPEG compression (MJPEG), Wavelet-based compression using the Snow codec (SNOW), H.264 videos subjected to simulated wireless transmission loss (Wireless), and Additive white noise (AWGN). However, in this work, we have only used the sequences with H264, Wireless, MJPEG, and HEVC distortions. All the videos are in the YUV420 format, with a spatial resolution of 832 $\times$ 480, a duration of 10 seconds, and temporal resolutions ranging from 24 to 60 fps. Each video was assessed by 35 participants following the SAMVIQ methodology. Figure 3.6 shows a frame of each video used in our tests.

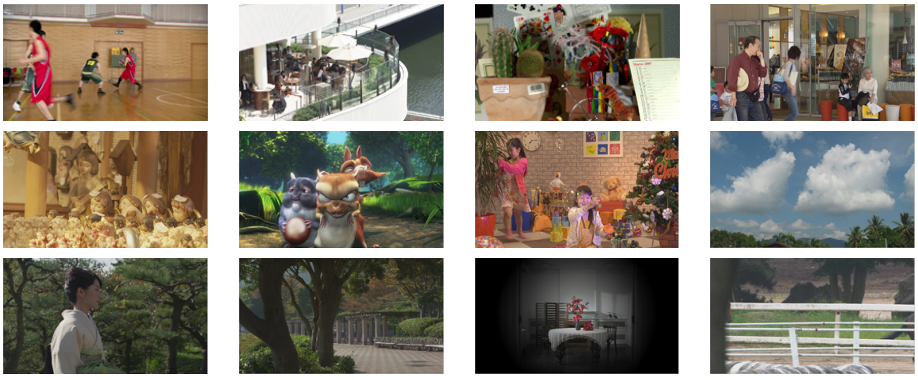


Figure 3.6: Sample images of source video contents from CSIQ database.

3.5 Statistical Analysis

Data gathered from the six experiments provided up to four different results for each test sequence: Eye-tracker data, Mean Annoyance Value (MAV), Mean Strength Value (MSV) and, Probability of detection (P_{det}). Also, the MSV values were divided in MSV_{bloc}, MSV_{blur}, and MSV_{pck}, which correspond to MSVs for blockiness, blurriness, and packet-loss, respectively.

To analyze the subjects' data gathered during detection tasks, we first converted the *yes/no* answers to binary scores, such as, *yes* was saved as 1, while *no* was saved as 0. The P_{det} an impairment is, then, estimated by counting the number of subjects who detect this impairment and dividing by the total number of subjects.

MAVs are computed by averaging the annoyance values over all participants, for each video:

$$MAV = \frac{1}{S}\sum_{i=1}^{S} A(i), \tag{3.1}$$

where $A(i)$ is the annoyance value reported by the i^{th} participant and, S is the number of human subjects.

MSVs are computed by averaging the strength values over all participants, for each video and artifact type:

$$MSV = \frac{1}{S}\sum_{i=1}^{S} S(i), \tag{3.2}$$

where $S(i)$ is the strength value reported by the i^{th} participant and, S is the number of human subjects. The MSV is computed for each type of artifact, i.e. blockiness, blurriness, or packed-loss. To study how the artifact strengths combine to predict the perceived annoyance of videos impaired by multiple and overlapping artifacts, we fit a set of linear and non-linear models to the MSV subjective data and the MAV data collected for the same test sequences [34, 79].

To estimate the performance of the models, we calculated the Pearson correlation coefficient (PCC) and the Spearman Rank Order Correlation Coefficient (SCC) between the subjective and predicted scores. Pearson coefficient measures the linear dependence (correlation) between two variables, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. Spearman coefficient measures the statistical dependence between the ranking of two variables. While Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships. When observation have a similar rank, the Spearman correlation between two variables is high (or identical for correlation of 1), and when observations have different rank it is high (or fully opposed for a correlation of -1).

We use the Akaike Information Criterion (AIC) [80] to analyze the trade-off between fitting accuracy and the number of degrees of freedom of the model, thereby controlling for the bias/variance trade-off and over-fitting. To test the effect of the artifact parameters on annoyance, we performed a Repeated-Measure ANOVA (RMANOVA) with a significance level of 95% ($\alpha = 0.05$).

Finally, we used a SVR algorithm to predict annoyance from the subjective data. The SVR algorithm is used to combine all *features* extracted from de videos impaired with blockiness, blurriness, and packet-loss. To train the SVR, we used a k-fold cross validation setup, i.e. we split the dataset in k equally sized non-overlapping sets and ran the training k times (with k equal 10). Each time, we use a different fold as test set, while using the remaining $k - 1$ folds for training [81].

3.6 Analysis of eye-tracking data and quality scores

To investigate the impact of the presence of multiple artifacts on users gaze patterns, we tracked eye-movements of participants during the psychophysical experiments (i.e. Exp.3a). Therefore, experiments produced two types of output: eye-tracking data and subjective scores. For both outputs, we had one recording per participant and per video content. The subjective scores collected for each video sequence were averaged over participants, as described in the previous section.

The eye-tracking data consisted of pupil movements, recorded in terms of fixation points and saccades. However, we limited ourselves to the analysis of fixation data, which is considered to be one of the most informative data regarding viewing behavior. Specifically, we analyzed viewing behavior by looking at two quantities: the duration of the fixations, found to be impacted by quality scoring in Ninassi *et al.* [65], and the spatial deployment of gaze patterns. We quantified the latter via saliency maps [82] which represent, per each video pixel, the probability that it will be gazed at by an average observer. This choice is in line with most of the existing literature in the area [61].

For each video, we recorded the fixation points on which the participant's pupil rests for at least 100ms. Then, we processed the fixation data to create the saliency maps [82] that were used to determine the most visually attractive areas in a scene. For videos, the fixation data is recorded at a frame level (i.e. every 20ms). However, calculating a saliency map for each video frame gives an excessive granularity, as compared to the duration of a fixation (around 400ms). Therefore, we adopted the same strategy used in our previous work [68] where for a given video, we grouped the fixations from all participants in time windows of 400ms, generating *fixation maps*. To compute the saliency map corresponding to a specific time window of a specific video, we smoothed the fixation maps by applying

a Gaussian patch with width equal to approximately the fovea size (2° visual angle). This procedure was applied for each video sequence, creating $\frac{10{,}000}{400}ms = 25$ saliency maps per video content. For the sake of analysis, the saliency maps are clustered into the following groups:

1. FV_{PV} corresponding to pristine videos, captured during the free-viewing task;
2. SC_{PV} corresponding to pristine videos, captured during the quality assessment task;
3. SC_{G1} corresponding to test sequences with artifacts in isolation (Combinations 2 to 4 in Table 3.3), captured during the quality assessment task;
4. SC_{G2} corresponding to test sequences with combinations of two artifacts (Combinations 5 to 12 in Table 3.3), captured during the quality assessment task;
5. SC_{G3} corresponding to test sequences with combinations of three artifacts (Combinations 13 to 20 in Table 3.3), captured during the quality assessment task.

Finally, we analysed if viewing behavior changes across these five groups from two independent variables: task (free-viewing or quality assessment) and degradation (pristine or impaired). For impaired videos, we also were interested in checking whether the number and/or type of artifacts impact the saliency maps and the fixation durations.

3.6.1 Similarity measures for detecting saliency changes

Similarity measures are used as indicators of changes in saliency distribution and therefore in gaze patterns [66, 83]. The following measures were adopted to estimate the extent to which saliency maps corresponding to a certain time window of a certain video changes across the groups indicated above.

1. Linear correlation coefficient (LCC) $\in$ [1 -1], which quantifies the strength of the linear relationship between two saliency maps.
2. Structure Similarity Index (SSIM [29]) $\in$ [0 1], which indicates the extent to which the structural information of a map is preserved in relation to another map.

In both similarity measures, a value close to 1 indicates high similarity, while a value of 0 indicates dissimilarity, in turn suggesting a consistent change in the image saliency and consequently of the spatial allocation of gaze.

We also used the measure Upper Empirical Similarity Limit (UESL) to account for content and inter-observer variability [66]. UESL represents the similarity of saliency maps obtained under the same experimental conditions (e.g. while observing the same

video, at the same level of impairment, and under the same task), but for two different groups of participants. As such, it expresses the extent to which two saliency maps are similar given individual differences in participants. UESL represents an useful benchmark to understand whether dissimilarity in maps, as measured after a change in experimental conditions (e.g. between a free-viewing map and a quality assessment map), is due to inter-subject variability rather than to the change in experimental conditions. We calculated UESL based on LCC using with the following equation:

$$UESL(LCC, v_i) = \frac{1}{T}\sum_{t=1}^{T} LCC\left(SM\left(v_{i,t}^{FV_{PV_0}}\right), SM\left(v_{i,t}^{FL_{PV_1}}\right)\right), \tag{3.3}$$

where $SM(v_{i,t}^{FL})$ indicates the saliency map computed for time slot t, video v_i, FV is the observer group, and T is the total number of fixations over all observers. The saliency maps $SM(v_{i,t}^{FV_{PV_0}})$ were recorded in our experiments from the pristine videos during the free-viewing task (FV_{PV}). The saliency maps $SM(v_{i,t}^{FV_{PV_1}})$ were also recorded from the pristine videos during the free-viewing task, but from a previous experiment [68]. To compute the UESL based on SSIM, we simply substitute LCC by SSIM in Eq. 3.3.

Chapter 4

Annoyance Models

Understanding the perceptual impact of compression artifacts in video is one of the keys for designing better coding schemes and appropriate visual quality control chains. Although compression and transmission artifacts, such as blockiness, blurriness and packet-loss, appear simultaneously in digital videos, traditionally they have been studied in isolation. In this chapter, we reported the analysis of the annoyance tasks performed in the three subjective experiments (i.e. Exp.1a, Exp.2a, and Exp.3a). Our goal here is to study the perceptual characteristics of a set of artifacts common in digital videos (blockiness, blurriness and packet-loss), presented in isolation and in combinations. Based on this analysis, we designed several annoyance models for videos degraded with these artifacts.

4.1 Introduction

Based on the results of three psychophysical experiments, we investigated how spatial and temporal artifacts combine to determine quality [68, 84] by measuring the annoyance and detection characteristics of two spatial artifacts (blockiness and blurriness) and a very important temporal artifact (packet-loss). Up to our knowledge, there is no study in the literature that performs an analysis of the influence spatial-temporal artifacts (in isolation and in combinations) have on the perceived annoyance. Most importantly, there is no study on how spatial and temporal artifacts interact to produce overall annoyance. To quantify the contribution of each artifact to the overall annoyance and of the interactions among the different artifacts, we tested linear and non-linear annoyance models.

4.2 Experiment 1: Packet-Loss

First, we analyzed the P_{det} for all test sequences of Exp.1a. Results showed that P_{det} increased with MAV. The P_{det} values for two of the videos were equal to one (videos

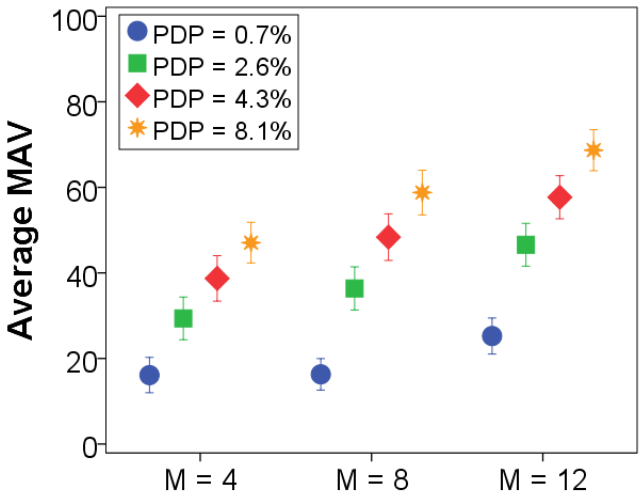


Figure 4.1: Exp.1a: Average MAV plots for different values of *PDP*: 0.7%, 2.6%, 4.3% and 8.1%.

Into Tree and *Barbecue*). This means that, for these two videos, all participants saw impairments in all test cases. It is worth pointing out that these two scenes have large smooth regions (e.g. skies) that make impairments easier to detect. *Park Joy*, *Cactus*, and *Basketball* have values of P_{det} that grow (and saturate) very fast as MSE increases. On the other hand, P_{det} for *Park Run* and *Romeo and Juliet* increases at a slower rate. This indicates that, for these scenes, it is harder to detect packet-loss artifacts. *Romeo and Juliet*, although having small spatial and temporal activity, is relatively dark and has a very clear focus of attention (the couple). On the other hand, *Park Run* has lots of spatial and temporal activity and not a lot of camera movement (see Figure 3.2). All of this makes it harder to spot packet-loss artifacts.

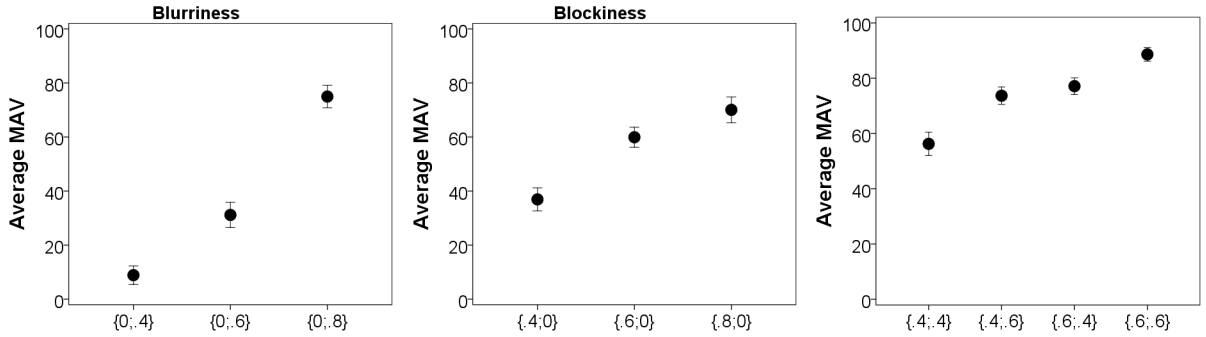
Next, we analyzed the influence of M and *PDP* on MAV. Figure 4.1 shows a plot of the average MAVs for the three values of M and the four values of *PDP*. Notice that MAV increases with both *PDP* and M, but *PDP* has a bigger effect on MAV than M. The effect of *PDP* on MAV is clearly significant. We performed a RM-ANOVA for analyzing the influence of M on MAV. Table 4.1 shows the pairwise comparisons between average MAVs for different M parameters. Notice that there are significant statistical differences between average MAVs for any pair of M values, except for the pair $M = 4$ and $M = 8$ for *PDP*=0.7%.

4.3 Experiment 2: Blockiness and Blurriness

We analyzed the P_{det} and we found values smaller than 0.20 for all original videos, except for *Into Tree* that has large smooth regions. Similarly to Exp.1a, test sequences with low

Table 4.1: Exp.1a: Pairwise comparisons between average MAVs for different M values. (* Significant at 0.05 level.)

M values		Diff. Mean	Std. Error
4	8	-0.170	1.512
4	12	-9.134*	1.664
8	12	-8.964*	1.946



(a) (b) (c)

Figure 4.2: Exp.2a: Average MAVs for: (a) blurriness, (b) blockiness, and (c) combinations of blockiness and blurriness.

P_{det} values got lower MAVs, while test sequences with higher P_{det} values got higher MAVs. Figures 4.2 (a) and (b) show plots of the average MAVs for sequences with only blockiness and blurriness, respectively, at strengths 0.4, 0.6, and 0.8. As expected, average MAVs increase with the artifact strength.

A RM-ANOVA was performed to check if MAV differences for different blockiness and blurriness strengths are significant. Table 4.2 displays the results, showing that there are significant statistical differences in MAV for all pairs of different strengths in only-blockiness and only-blurriness sequences.

Table 4.2: Exp.2a: Pairwise comparisons of MAVs for videos with only blockiness ($\hat{F} = 85.62, \alpha \leq 0.01$) and only blurriness ($\hat{F} = 334.75, \alpha \leq 0.01$). (* Significant at 0.05 level)

		Blockiness		Blurriness	
Strengths		Diff. Mean	Std. Error	Diff. Mean	Std. Error
0.4	0.6	-22.982*	1.863	-22.295*	2.796
	0.8	-33.125*	3.179	-66.107*	2.526
0.6	0.8	-10.143*	2.571	-43.813*	2.464

Table 4.3: Exp.2a: Pairwise comparisons of MAVs of sequences with combinations of blockiness and blurriness. (* Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(0.4;0.4)	(0.4;0.6)	-17.420*	2.044
	(0.6;0.4)	-20.866*	1.841
	(0.6;0.6)	-32.375*	2.044
(0.4;0.6)	(0.6;0.4)	-3.446	1.499
	(0.6;0.6)	-14.955*	1.445
(0.6;0.4)	(0.6;0.6)	-11.509*	1.097

Figure 4.2(c) shows a plot of the average MAVs for sequences of blockiness and blurriness in combination. A RM-ANOVA was performed to test pairwise comparisons of the average MAVs of these sequences (see Table 4.3). Results show that there are significant statistical differences between average MAVs obtained for any pair of blockiness and blurriness combinations ($\hat{F} = 124.68, \alpha \leq 0.01$), except for the pair (0.4;0.6) and (0.6;0.4). This means that a change in the artifact strength was perceived by human subjects.

We also checked if there were differences between sequences with one and two artifacts, for example, only blockiness (0.4;0.0) with blockiness and blurriness in combination (0.4;0.4). Results of the pairwise comparisons are showed in Tables 4.4 and 4.5. Notice that significant statistical differences between average MAVs for any pair of combinations were found. In other words, on average, adding an extra artifact affected the MAV.

Table 4.4: Exp.2a: Pairwise comparisons of MAVs between sequences with only artifacts and sequences with combinations of artifacts (* Significant at 0.05 level).

Combinations		Diff. Mean	Std. Error
(0.4;0.0)	(0.4;0.4)	-19.330*	2.027
	(0.4;0.6)	-36.750*	2.453
(0.6;0.0)	(0.6;0.4)	-17.214*	1.844
	(0.6;0.6)	-28.723*	1.921

Table 4.5: Exp.2a: Pairwise comparisons of MAVs between sequences with only blurriness and sequences with combinations of blockiness and blurriness (* Significant at 0.05 level).

Combinations		Diff. Mean	Std. Error
(0.0;0.4)	(0.4;0.4)	-47.393*	2.492
	(0.6;0.4)	-68.259*	2.124
(0.0;0.6)	(0.4;0.6)	-42.518*	2.507
	(0.6;0.6)	-57.473*	2.553

4.4 Experiment 3: Blockiness, Blurriness and Packet-Loss

We analyzed the P_{det} for all original videos and we found values below 0.09, except for *Park Run* video ($P_{det} = 0.17$). *Park Run* has a lot of spatial and temporal activity and not a lot of camera movement, what could have led some participants to think they saw impairments in the originals. Similarly to Exp.1a and Exp.2a, test sequences with low P_{det} values got lower MAVs, while test sequences with higher P_{det} values got higher MAVs.

Figure 4.3 show plots of average MAV over all test sequences with pure strong blockiness (0.0;0.6;0.0), blurriness (0.0;0.0;0.6), and packet-loss (8.1;0.0;0.0). For comparison purposes, the plot also shows the average MAVs for the original sequences. As expected, the average MAV for originals is close to zero and, when the artifact is added at increasing values, MAV also increases. Average MAV values are higher for blockiness (average MAV for bloc= 0.6 is 48.56), followed by packet-loss (average MAV for PDP=8.1% is 37.99), and blurriness (average MAV for blur=0.6 is 32.45). This is in agreement with results of Exp.2a, where blockiness artifacts are the most annoying artifacts. To check if these average MAVs differences between artifacts were statistically significant, we performed a RM-ANOVA (Table 4.6) that found MAV differences between blockiness and the other two artifacts were significant ($\hat{F} = 24.906, \alpha \leq 0.01$). However, the difference in average MAVs between packet-loss and blurriness were not statistically significant.

Table 4.6: Exp.3a: Pairwise comparisons for sequences with only packet-loss, blockiness and blurriness. (*. Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(8.1;0.0;0.0)	(0.0;0.6;0.0)	-10.590*	2.006
	(0.0;0.0;0.6)	5.534	2.701
(0.0;0.6;0.0)	(0.0;0.0;0.6)	16.124*	2.203

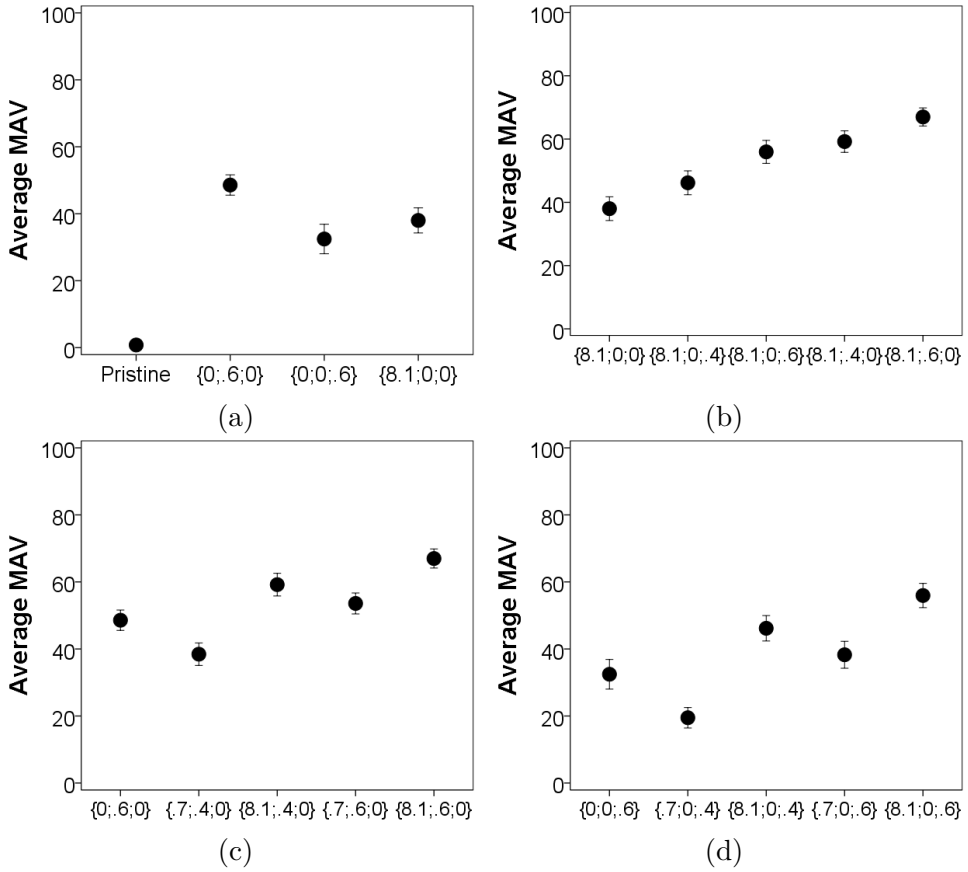


Figure 4.3: Exp.3a: (a) Average MAVs for blockiness, blurriness and packet-loss, (b) MAVs for packet-loss by itself (PDP) and in combination with blurriness (+blur) and blockiness (+bloc), (c) MAVs for blockiness by itself (bloc) and in combination with packet-loss ($+PDP$), and (d) MAVs for blurriness by itself (blur) and in combination with packet-loss ($+PDP$).

Figure 4.3 (b) shows a plot of the average MAV for test sequences with combinations of strong packet-loss artifacts (PDP=8.1%) and either blockiness (bloc=0.4 or 0.6) or blurriness (blur=0.4 or 0.6). Table 4.7 shows the RM-ANOVA test performed on the average MAVs of these sequences. Results of a RM-ANOVA pairwise comparisons found significant statistical differences between the average MAVs obtained for combinations of packet-loss and either blockiness or blurriness artifacts ($\hat{F} = 99.542, \alpha \leq 0.01$), except for the pair of combinations (8.1;0.0;0.6) and (8.1;0.4;0.0), indicating that combining packet-loss with either blockiness and blurriness, on average, affects the MAV.

Figure 4.3 (c) shows a plot of the average MAV for test sequences with packet-loss and blockiness artifacts. Table 4.8 shows the results of the RM-ANOVA tests performed on the average MAVs of these sequences. Notice that there are significant statistical differences for all pairs of combinations ($\hat{F} = 101.252, \alpha \leq 0.01$). Figure 4.3 (d) shows a plot of the average MAV for test sequences with combinations packet-loss and blurriness.

Table 4.7: Exp.3a: Pairwise comparisons for sequences with packet-loss and either blockiness or blurriness. (*. Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(8.1;0.0;0.0)	(8.1;0.0;0.4)	-8.180*	1.624
	(8.1;0.0;0.6)	-17.950*	1.838
	(8.1;0.4;0.0)	-21.199*	1.509
	(8.1;0.6;0.0)	-28.994*	1.653
(8.1;0.0;0.4)	(8.1;0.0;0.6)	-9.770*	1.659
	(8.1;0.4;0.0)	-13.019*	1.668
	(8.1;0.6;0.0)	-20.814*	1.620
(8.1;0.0;0.6)	(8.1;0.4;0.0)	-3.248	1.555
	(8.1;0.6;0.0)	-11.043*	1.488
(8.1;0.4;0.0)	(8.1;0.6;0.0)	-7.795*	1.418

Again, a RM-ANOVA test (Table 4.9) found significant statistical differences for all pairs of these combinations ($\hat{F} = 93.310, \alpha \leq 0.01$). In general, combinations of packet-loss and blockiness have higher average MAVs than combinations of packet-loss and blurriness. Also, for combinations of packet-loss, blockiness, and blurriness, the presence of an additional artifact incurs in an increase of the average MAVs.

Table 4.8: Exp.3a: Pairwise comparisons for sequences with combinations of packet-loss and blockiness artifacts. (*. Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(0.0;0.6;0.0)	(0.7;0.4;0.0)	10.137*	1.554
	(8.1;0.4;0.0)	-10.609*	1.787
	(0.7;0.6;0.0)	-4.994*	1.343
	(8.1;0.6;0.0)	-18.404*	1.536
(0.7;0.4;0.0)	(8.1;0.4;0.0)	-20.745*	1.677
	(0.7;0.6;0.0)	-15.130*	1.439
	(8.1;0.6;0.0)	-28.540*	1.494
(8.1;0.4;0.0)	(0.7;0.6;0.0)	5.615*	1.506
	(8.1;0.6;0.0)	-7.795*	1.418
(0.7;0.6;0.0)	(8.1;0.6;0.0)	-13.410*	1.343

Table 4.9: Exp.3a: Pairwise comparisons for sequences with combinations of blurriness and packet-loss artifacts. (*. Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(0.0;0.0;0.6)	(0.7;0.0;0.4)	12.975*	2.310
	(8.1;0.0;0.4)	-13.714*	2.732
	(0.7;0.0;0.6)	-5.820*	1.749
	(8.1;0.0;0.6)	-23.484*	2.122
(0.7;0.0;0.4)	(8.1;0.0;0.4)	-26.689*	1.812
	(0.7;0.0;0.6)	-18.795*	1.983
	(8.1;0.0;0.6)	-36.460*	1.756
(8.1;0.0;0.4)	(0.7;0.0;0.6)	7.894*	2.177
	(8.1;0.0;0.6)	-9.770*	1.659
(0.7;0.0;0.6)	(8.1;0.0;0.6)	-17.665*	1.625

4.5 Comparison of Data from Experiments

Research shows that even results gathered from experiments using the same experimental methodology may differ considerably because of differences in physical location, viewer expectations, and especially set of stimuli [12]. It is known that participants have a tendency to use the entire scoring scale to evaluate the quality of the test stimuli presented in an experimental session. As consequence, scores may suffer from context effects [85]. For example, mildly impaired stimuli may get higher annoyance scores in an experiment containing only unimpaired or slightly impaired stimuli than in an experiment containing slightly to highly impaired stimuli.

In our experiments, we used different artifacts at different strengths. It is reasonable to assume that they may have spanned different ranges of MAVs that are not necessarily equivalent. In other words, the highest MAVs in the three experiments may correspond to videos impaired with artifacts of very different perceptual strengths. For example, videos with the highest packet-loss strengths in Exp.1a may have received the highest MAVs. But, the same MAVs in Exp.3a may correspond to videos with much more annoying artifacts (and a lower quality), most likely presenting packet-loss in combination with blockiness and blurriness.

In fact, if we compare the MAVs obtained by sequences with strongest packet-loss configuration in isolation (i.e. (8.1;0;0)) in Exp.1a and Exp.3a, we see a striking difference. In Exp.1a, this is the highest level of impairment encountered by participants throughout the whole experiment. As such, it obtains a relatively high MAV (on average, across

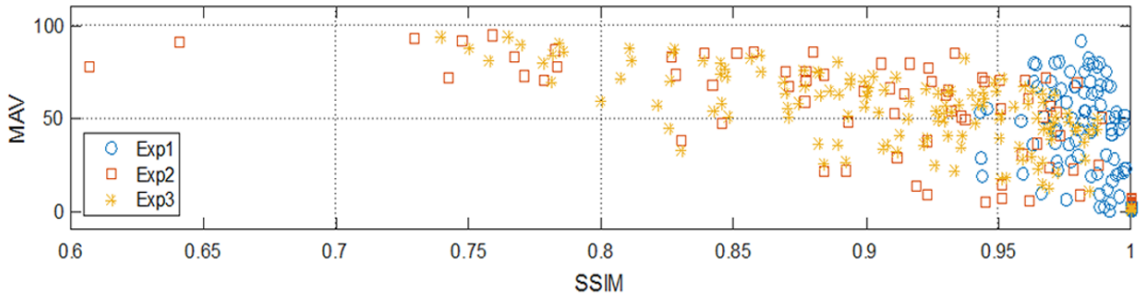
all contents, MAV > 70, see Figure 4.1). On the other hand, the same videos impaired with the same combination in Exp.3a, are perceived only as mildly annoying (across all contents, $MAV \sim 40$, see Figure 4.3 (a)). This is probably because, in comparison with videos that are distorted by multiple artifacts, heavy packet-loss is not as annoying. This discrepancy clearly points towards the presence of context effects in the MAVs of Exp.1a: MAVs are artificially inflated due to the relatively narrow range of quality spanned by the videos included in experiment. A re-alignment process is therefore necessary to map the MAVs of Exp.1a to a range that is more commensurate to the annoyance values measured in Exp.2a and Exp.3a.

Pinson *et al.* proposed a technique to merge data from different experiments known as the Iterative Nested Least Squares Algorithm (INLSA) [12,86]. INLSA re-scales subjective scores from different experiments using objective quality metrics as a common external variable. The procedure is performed solving two least squares problems. A single first-order correction method is used in the first problem to homogenize the heterogeneous scores of the different experiments. An approximation of the linear combinations of the parameters across the scores of the different experiments is obtained by solving the second problem. A full mapping of the scores of the different experiments into a common scale is obtained by performing an iteration of these two least-squares problems. To sample the mapping among scores of the different experiments, it is necessary to choose a common set of stimuli from all the experiments involved in the realignment.

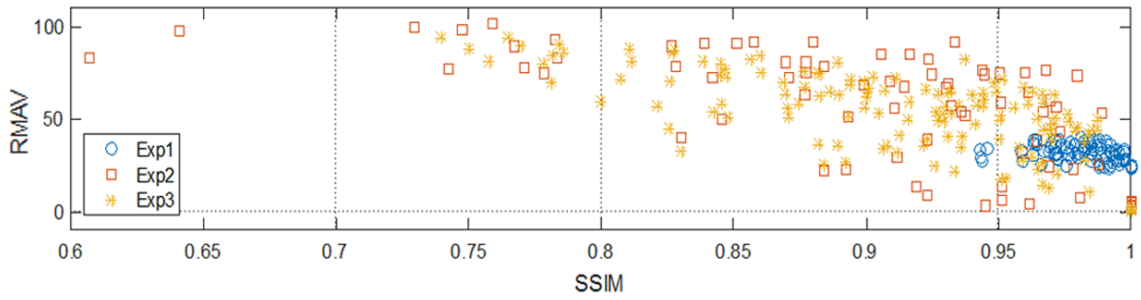
Before comparing the data of the three experiments, we used INLSA to re-align the annoyance scores. We used SSIM [29] as the objective quality metric. Exp.3a was used as the reference experiment because it had the highest number of artifact combination. Figures 4.4 show the MAV for the complete set of experiments before (top) and after (bottom) using INLSA, respectively, against the corresponding SSIM values [29] value of the video. Notice that for the same SSIM values each experiment has a different range of MAVs. In particular, and as expected, for Exp.1a, the entire MAV range is clustered on the top part of the SSIM scale. This means that videos with relatively low levels of impairments (as measured by SSIM) are judged as highly annoying (probably due to context effects, as mentioned above). This is not true for the other two experiments.

After mapping the MAVs from Exp.1a and Exp.2a into the scale of Exp.3a, the MAVs of Exp.1a span a more comparable range of annoyance. The range of the RMAVs of Exp.1a ($Avg. = 32.17$, $Std = 5.52$) is smaller than the original range spanned by its MAVS ($Avg. = 42.78$, $Std. = 25.47$) and more skewed towards the lower part of the annoyance scale. In other words, RMAVs now denote that annoyance values of videos impaired with only packet-loss (as it is the case for Exp.1a) are lower when compared with those of sequences distorted by multiple artifacts. This result suggests that scores

from Exp.1a can be merged with those of the other two experiments, making it possible to analyze the data from the three experiments as a whole.



(a)



(b)

Figure 4.4: (a) MAVs and (b) RMAVs (after applying INLSA [86]) versus SSIM for Exp.1a, Exp.2a, and Exp.3a.

4.5.1 Annoyance Models

Aiming to study if models that combine the artifact strength values (*PDP*, *bloc*, and *blur*) can predict the perceived annoyance of videos impaired by multiple and overlapping artifacts, we fitted a set of linear and non-linear models. Prior to fitting the models, all artifact strength values were normalized to the same range $[0, 1]$, with the strength value *1* corresponding to the strongest artifact level found in practice and the value *0* corresponding to the absence of the artifact. Blockiness and blurriness were already generated using this scale, but the packet-loss strength values needed to be normalized. To re-scale *PDP* to fit this range, we assumed that *PDP* values greater than 10 would be unrealistic in practical network conditions and set 10 as the maximum *PDP* value [87,88]. The normalized packet-loss strength is obtained by dividing the original values by 10, i.e. $pdp = PDP/10$.

Linear Models

The first linear model we tested was a simple linear model without any interaction term, given by:

$$PA_{L1} = \alpha \cdot \text{pdp} + \beta \cdot \text{bloc} + \gamma \cdot \text{blur}, \tag{4.1}$$

where PA_{L1} corresponds to the predicted (non-realigned) MAVs and *pdp*, *bloc*, and *blur* correspond to the strength of each artifact. Line 2 of Table 4.10 shows the results of the fitting. We also adapted Equation 4.1 to include an intercept coefficient (δ), referring to this model as PA_{L2}:

$$PA_{L2} = \alpha \cdot \text{pdp} + \beta \cdot \text{bloc} + \gamma \cdot \text{blur} + \delta. \tag{4.2}$$

Line 3 of Table 4.10 shows the fit results for MAVs prior to the re-alignment with INLSA.

We tested the above models (Equations 4.1 and 4.2) on the MAVs re-aligned using INLSA, hereafter referred to as RMAVs. Line 4 of Table 4.10 shows the results for the first linear model (PRA_{L1}) fit, while line 5 shows the results for the second linear model (PRA_{L2}, with intercept term). To evaluate the goodness of the fit of each model, we report the Pearson correlation coefficient (PCC) and Spearman correlation coefficient (SCC) between predicted and subjective MAVs (or RMAVs) where the fit was based on the entire dataset. For both models, a better fit was obtained using RMAV instead of MAV.

Table 4.10: Fitting of the linear models to MAV and RMAV.

Models	δ	α	β	γ	PCC	SCC
PA_{L1}		50.060	72.480	48.620	0.726	0.721
PA_{L2}	23.515	29.350	50.606	26.740	0.730	0.727
PRA_{L1}		35.770	78.404	52.602	0.844	0.867
PRA_{L2}	18.170	19.768	61.499	35.698	**0.850**	**0.870**

Linear Models with Interactions

It has be shown that interaction terms must be taken into account when modeling the annoyance caused by combinations of artifacts because masking and facilitation processes may occur when artifacts are combined [10]. To investigate if the presence of one artifact may affect the perception of the other(s) and how this impacts the overall annoyance, we

Table 4.11: Fitting of the linear model with interactions (PA_{L3}) to MAVs.

Coef.	Estimate	Std. Error	t-value	$Pr(>\|t\|)$
α	85.024	3.302	25.749	$< 2e-16^a$
β	88.550	4.344	20.386	$< 2e-16^a$
γ	64.118	4.344	14.761	$< 2e-16^a$
ρ_1	-123.393	13.301	-9.277	$< 2e-16^a$
ρ_2	-120.320	13.301	-9.046	$< 2e-16^a$
ρ_3	-22.561	14.724	-1.532	0.127
ρ_4	175.670	38.860	4.521	$< 8.87e{-}06^a$
[a] Statistically significant at ($P < 0.05$)			PCC = 0.860, SCC = 0.841.	

fitted a linear model with interactions, (PA_{L3}), defined as:

$$\begin{aligned} PA_{L3} = \alpha \cdot \text{pdp} + \beta \cdot \text{bloc} + \gamma \cdot \text{blur} + \\ \rho_1 \cdot \text{pdp} \cdot \text{bloc} + \rho_2 \cdot \text{pdp} \cdot \text{blur} + \\ \rho_3 \cdot \text{bloc} \cdot \text{blur} + \rho_4 \cdot \text{pdp} \cdot \text{bloc} \cdot \text{blur}. \end{aligned} \tag{4.3}$$

Results of this fit for MAVs are shown in Table 4.11. Column 2 of this table shows the values of the model coefficients, while column 5 shows the corresponding p-values (based on t-test, two-tailed, $p < 0.05$). Notice that the first, second, and third order coefficients are statistically significant, with the exception of for the ρ_3 coefficient corresponding to the interaction of blockiness and blurriness.

We also tested the same model with the addition of an intercept term δ, denoted as PA_{L4}. The results of this fit for MAVs are shown in Table 4.12. Again, the first, second, and third order terms have a statistically significant effect, with the exception of ρ_3 coefficient that corresponds to the interaction of blockiness and blurriness.

Fitting the two linear models with interaction terms with and without a fixed intercept on RMAVs, we obtained the predictions (PRA_{L3}) and (PRA_{L4}). Tables 4.13 and 4.14 show the results obtained for both model without the intercept coefficient (PRA_{L3}) and for the model with the intercept coefficient (PRA_{L4}), respectively. For both models, all main effects and first order interactions are statistically significant, except for ρ_3 (interaction of blockiness and blurriness) in PRA_{L3}. The second order interactions are not statistically significant for both models. Correlation coefficients are higher when RMAVs are used.

Table 4.12: Fitting of the linear model with interactions (PA_{L4}) to MAVs.

Coef.	Estimate	Std. Error	t-value	$Pr(>\|t\|)$
δ	14.117	2.078	6.792	$5.95e-11^a$
α	62.207	4.557	13.650	$<2e-16^a$
β	65.050	5.327	12.211	$<2e-16^a$
γ	40.619	5.327	7.625	$3.24e-13^a$
ρ_1	-84.372	13.670	-6.172	$2.18e-09^a$
ρ_2	-81.299	13.670	-5.947	$7.58e-09^a$
ρ_3	15.613	14.836	1.052	0.29348
ρ_4	109.970	37.507	2.932	0.00363^a
[a] Statistically significant at ($P < 0.05$)			PCC = 0.853, SCC = 0.823.	

Table 4.13: Fitting of the linear model with interactions (PRA_{L3}) for RMAVs.

Coef.	Estimate	Std. Error	t-value	$Pr(>\|t\|)$
α	57.064	2.784	20.494	$<2e-16^a$
β	88.685	3.663	24.212	$<2e-16^a$
γ	61.703	3.663	16.846	$<2e-16^a$
ρ_1	-69.785	11.217	-6.222	$<1.65e-09^a$
ρ_2	-63.363	11.217	-5.649	$<3.74e-08^a$
ρ_3	-10.196	12.416	-0.821	0.4122
ρ_4	55.827	32.768	1.704	0.0895
[a] Statistically significant at ($P < 0.05$)			PCC = 0.880, SCC = 0.886.	

Non-Linear Models

The proposed linear models, although fairly accurate, may be unable to capture the complex non-linear interactions of the artifact combinations [89]. Therefore, we tested two different types of non-linear models: a Minkowski metric model and a model based on SVR. We tested two Minkowski metrics, one without the intercept term (PA_{M1}) and another with the intercept term (PA_{M2}), as given by the following equations:

$$PA_{M1} = (\text{pdp}^m + \text{bloc}^m + \text{blu}^m)^{\frac{1}{m}}, \tag{4.4}$$

and

$$PA_{M2} = (\delta + \text{pdp}^m + \text{bloc}^m + \text{blu}^m)^{\frac{1}{m}}, \tag{4.5}$$

Table 4.14: Fitting of the linear model with interactions and with an intercept coefficient (PRA_{L4}) for RMAVs.

Coef.	Estimate	Std. Error	t-value	$Pr(>\|t\|)$
δ	14.420	1.689	8.540	$6.83e-16^a$
α	33.757	3.702	9.118	$<2e-16^a$
β	64.681	4.328	14.946	$<2e-16^a$
γ	37.698	4.328	8.711	$<2e-16^a$
ρ_1	-29.924	11.105	-2.695	0.00744^a
ρ_2	-23.503	11.105	-2.116	0.03514^a
ρ_3	28.800	12.053	2.390	0.01749^a
ρ_4	-11.286	30.470	-0.370	0.71134
[a] Statistically significant at ($P < 0.05$)			PCC = 0.871, SCC = 0.886.	

where PA_{M1} and PA_{M2} are the predicted annoyance, and m is the Minkowski power, obtained as a result of the fitting. It is worth pointing that this is the same combination rule used by Huib de Ridder [18] and Farias *et al.* [10] to predict annoyance caused by blockiness, blurriness, noisiness, and ringing. De Ridder's model was tested on a smaller data set of still images and returned $m > 1.6$ values, whilst Farias's model was tested on interlaced SD videos and returned $m > 0.8$ values. Our results are different from the results obtained by both authors, what is expected since our stimuli consist of HD videos with both spatial and temporal artifacts.

Lines 2 and 3 of Table 4.15 show the results of the fit on non-realigned MAVs of the model without intercept term (PA_{M1}) and the model with intercept (PA_{M2}), respectively. Lines 4 and 5 show the results of fitting on re-aligned MAVs of the model without intercept term (PRA_{M1}) and the model with intercept (PRA_{M2}), respectively. We can observe that these non-linear models perform worse than the linear ones. Within non-linear models, we observe again a better performance of those fit on RMAVs.

Table 4.15: Fitting of Minkowski models on MAV and RMAV.

Models	m	δ	PCC	SCC
PA_{M1}	0.215		0.472	0.652
PA_{M2}	0.419	4.018	0.660	0.654
PRA_{M1}	0.215		0.562	**0.770**
PRA_{M2}	0.397	3.424	**0.770**	0.744

Finally, we used SVR to predict annoyance from the artifact strength data using both MAVs and RMAVs. Machine learning-based approaches such as SVR have been shown to be suitable to model complex non-linear perceptual processes related to artifact annoyance [89]. In these approaches, the model is not previously defined but is learned from the data (i.e. our database of impaired videos). To train SVR, we used a k-fold cross validation setup. We split the dataset in k equally sized, non-overlapping sets. We then ran the training k times, for each of which a different fold was used as test set, and the remaining $k-1$ folds were used for training. In this way, each data point has a chance of being validated against the other. In our experiments, we set k to 10, thereby running 10 repetitions of the training. We then computed the correlation between subjective data and model predictions per each run, and took their average as the SVR model performance measure. The SVR trained on RMAVs returned PCC and SCC values equal to 0.855 and 0.833, respectively, whereas the model trained on MAVs returned PCC and SCC values equal to 0.850 and 0.828, respectively.

Model comparison

The different models considered in the previous session achieved different degrees of accuracy, yet in some cases at the expense of increased complexity. For example, models with interaction terms have more degrees of freedom (i.e. parameters to be fit) than models without; as a consequence, although more accurate, they may be more prone to overfitting. To compare the models in terms of the trade-off between complexity and accuracy, we used the Akaike Information Criterion (AIC) [80]. AIC expresses the trade-off between accuracy of fitting and the number of degrees of freedom in the model, thereby controlling for the bias/variance trade-off and overfitting. Table 4.16 summarizes the AIC values computed considering the tested models, where a model with lower AIC is preferred. Notice that although PRA_{L4} (the linear model with interaction and bias terms fit on re-aligned data) has more parameters, it has the lowest AIC, i.e. the best trade-off between goodness-of-fit and complexity.

To verify whether the PRA_{L4} model also gives the best performance in terms of correlation, we performed again its fitting and that of all the other models in a 10-fold cross-validation setting, to obtain measurements comparable to those obtained for the SVR. The outcomes are reported in Table 4.17. Notice that PRA_{L4} outperforms all models, including SVR.

Table 4.16: Akaike Information Criterion for the linear and Minkowski models. A lower value indicates a better trade-off between model complexity and accuracy.

Model	df	AIC	Model	df	AIC
PA_{L1}	4	2776.212	PA_{L4}	9	2562.162
PRA_{L1}	4	2607.776	PRA_{L4}	9	**2434.164**
PA_{L2}	5	2638.636	PA_{M1}	2	3207.925
PRA_{L2}	5	2464.547	PRA_{M1}	2	3144.523
PA_{L3}	8	2604.215	PA_{M2}	3	2693.669
PRA_{L3}	8	2499.193	PRA_{M2}	3	2608.433

Table 4.17: Average correlation across the 10-fold cross-validation runs between model predictions and (R)MAVs

Model	PCC	SCC	Model	PCC	SCC
PA_{L1}	0.706	0.713	PA_{M1}	0.463	0.628
PRA_{L1}	0.836	0.849	PRA_{M1}	0.560	0.745
PA_{L2}	0.711	0.719	PA_{M2}	0.640	0.630
PRA_{L2}	0.844	0.851	PRA_{Mn2}	0.736	0.745
PA_{L3}	0.775	0.747	PA_{SVM}	0.855	0.834
PRA_{L3}	0.849	**0.867**	PRA_{SVM}	0.851	0.829
PA_{L4}	0.782	0.762			
PRA_{L4}	**0.861**	0.858			

4.6 Discussion

Models fit on RMAVs obtained a better performance, showing that re-aligning the data before fitting the models is beneficial. When an intercept constant was added to the models, the correlation coefficients increased. One possible cause for this result is that the original content may contain pre-existing artifacts, which participants judged as slightly annoying.

For all linear models, the coefficients corresponding to *bloc* had the highest magnitude, indicating that blockiness had the biggest impact on the perceived annoyance. When fitting linear models on MAVs, *pdp* had a stronger impact on annoyance than *blur*, while when the fitting was done on RMAVs, *blur* had a higher impact. This divergence is caused by the fact that MAVs corresponding to sequences affected by packet-loss in Exp.1a were overestimated (probably due to context effects). Therefore, when no re-alignment was

performed, the exaggerated MAVs of sequences with packet-loss caused this artifact to have a higher impact.

The majority of the second order coefficients were statistically significant. For the models fitted on MAVs, the exception is ρ_3, indicating that the specific combination of blockiness and blurriness does not influence the annoyance scores. In fact, for models fitted on MAVs, the majority of the interaction coefficients that include *pdp* were statistically significant. For models fit on RMAVs, the ρ_3 in the PRA_{L3} model (without intercept) was also not statistically significant. Most second order coefficients were negative, implying that the overall annoyance caused by the presence of two artifacts is not simply an addition of the respective annoyances. The co-presence of two artifacts might, in fact, reduce their combined overall annoyance. In other words, there may be masking effects among artifacts, with artifacts mutually attenuating each other's strength. Interaction coefficients with higher magnitudes were those corresponding to the *pdp·bloc* and *pdp·blur* terms. This suggests that packet-loss affects how blockiness and blurriness are perceived.

Third order interaction coefficients (ρ_4) were significant for MAVs and non-significant for RMAVs. Again, since in the non-realigned MAV set the contribution of the *pdp* parameter was overestimated, any interaction term containing *pdp* (ρ_1, ρ_2, and ρ_4) had a statistically significant impact on MAVs. This is not true for models fit on RMAVs, for which the specific strength combination of the three artifacts did not contribute to the overall annoyance.

Correlation coefficients obtained for Minkowski models were lower than what was obtained for the linear models. The Minkowski powers found ($0.215 < m < 0.420$) were considerably lower than the values found by other authors [10, 18]. This may indicate that these models were, in fact, more sensitive to small changes in artifact strengths. For these models we obtained similar correlation coefficients for the fits on MAVs and RMAVs. Finally, the SVR-based approach achieved correlations slightly lower than those achieved by the best linear model PRA_{L4}. Therefore, in this setting, linear models have a better accuracy performance.

Chapter 5

Strength Models

In this chapter, we presented a study of the characteristics and relationship between the perceptual strengths of blockiness, blurriness, and packet-loss artifacts. Similar to what was done in the previous chapter, we also proposed models that show how perceptual strengths can be combined to estimate the overall annoyance.

5.1 Introduction

When a video is degraded by the presence of several types of artifacts, the perceived quality is affected [17,18,90–92]. Therefore, alternatives to regular quality metrics include artifact metrics [15, 51] that measure the strength of individual artifacts. Given that the overall video quality can be estimated by combining the individual artifact *perceptual* strengths, the output of these metrics can be combined to obtain an overall annoyance score [4]. Naturally, there is a considerable number of no-reference metrics that uses this *multidimensional* approach for measuring the overall quality of a video [16, 31, 52].

Nevertheless, the performance of an artifact-based metric depends on the performance of the individual artifact metrics. Aiming to analyze the relationship between the perceptual strengths of blockiness, blurriness, and packet-loss artifacts, and how they can be combined to estimate the overall annoyance, we analyzed the results of the strength task gathered during the three psychophysical experiments performed in this work. In the next sections, we present the data analysis for Exp.1s, Exp.2s, and Exp.3s.

5.2 Experiment 1: Packet-Loss

In Exp.1s, participants rated the strength of test sequences impaired with only packet-loss. Figure 5.1 (a) shows a graph of the average MSV_{pck} versus PDP, grouped according to the M value. We also calculated the average MSV_{pck} for the original videos (blue ball

Table 5.1: Exp.1s: Pairwise comparisons between average MSV_{pck} with different PDP values for $M = 12$. (* Significant at 0.05 level.)

PDP values		Diff. Mean	Std. Error
0.7	2.6	-21.541*	2.434
0.7	4.3	-29.684*	2.280
0.7	8.1	-35.918*	3.077
2.6	4.3	-8.143*	2.357
2.6	8.1	-14.378*	2.584
4.3	8.1	-6.235	2.492

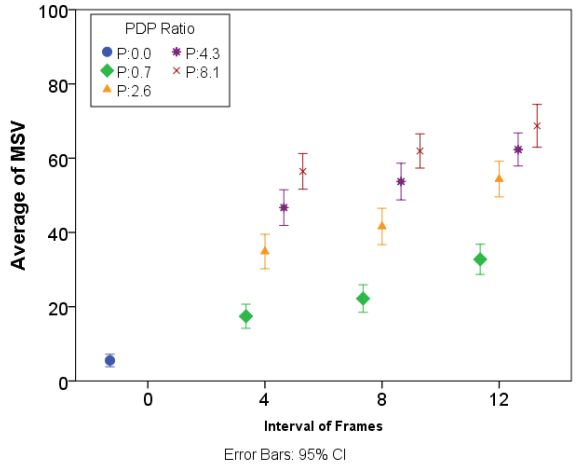


Figure 5.1: Exp.1s: MSV_{pck} plots for clustered error for $M = 4$,8, and 12.

on left-bottom). We can notice that the MSV_{pck} values are not equal to zero, indicating that participants perceived impairments in unimpaired videos. For $M = 4$, 8, and 12, the highest MSV_{pck} always corresponded to the strongest artifact (i.e. $PDP = 8.1\%$). Although MSV_{pck} increases with both PDP and M, PDP seems to have a bigger effect on MSV_{pck} than M.

A RM-ANOVA was performed to check the influence of the parameters M and the PDP on the MSV_{pck}. Results showed that there were significant statistical differences between the average of MSV_{pck} obtained for any pair of M values. When we analyze the influence of PDP on MSV_{pck}, we verified that there are significant statistical differences between the MSV_{pck} values for most PDP pairs, except for $PDP = 4.3\%$ and $PDP = 8.1\%$ for $M = 12$ (see Table 5.1).

Table 5.2: Exp.1s: Fitting parameters for linear model without intercept ($PA_{E1,L1}$) (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($>\|t\|$)	PCC	SCC
α	0.904	0.016	56.150	$< 2e-16*$	0.953	0.949

Table 5.3: Exp.1s: Fitting parameters for linear model with intercept ($PA_{E1,L2}$). (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($>\|t\|$)	PCC	SCC
δ	-4.396	1.607	-2.736	0.007*	0.953	0.950
α	0.983	0.033	29.816	$< 2e-16*$		

Table 5.4: Pearson and Spearman correlation coefficient of the linear models with and without intercept term, and SVR models on MAV.

Models	PCC	SCC
$PA_{E1,L1}$	0.953	0.949
$PA_{E1,L2}$	0.953	0.950
$PA_{E1,SVR}$	0.953	0.927

With the goal of studying if MSV can predict the perceived annoyance, we tested the following simple linear model without any interaction term:

$$PA_{E1,L1} = \alpha \cdot MSV_{pck}, \tag{5.1}$$

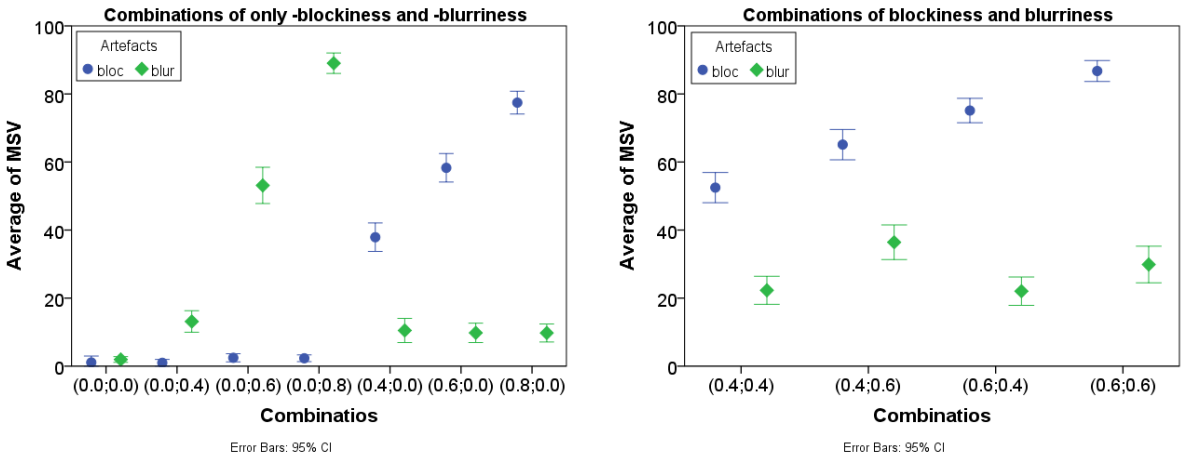
and the model with an intercept term δ:

$$PA_{E1,L2} = \delta + \alpha \cdot MSV_{pck}. \tag{5.2}$$

The outcomes of both linear models are depicted in Tables 5.2 and 5.3, where we can notice all coefficients are statistically significant. We also used SVR to predict annoyance from the strength data using MSV_{pck}. We refer to this model as $PA_{E1,SVR}$. PCC and SCC values obtained from the trained SVR were 0.953 and 0.927, respectively. Table 5.4 presents a summary of fitting for both linear model with and without intercept term, and SVR model. As we can notice, all model are similar performance considering the correlation coefficient.

5.3 Experiment 2: Blockiness and Blurriness

As mentioned earlier, test sequences used in Exp.2s had two different types of artifacts: blockiness and blurriness. These artifacts were presented in different strengths, either in isolation or in combination. Figure 5.2 (a) shows a graph of the average MSV_{blur} (green) and the average MSV_{bloc} (blue) for test sequences containing combinations of only-blurriness and only-blockiness. The first combination of the graph corresponds to the pristine videos. Notice that, again, the MSVs are not equal to zero, indicating that participants perceived impairments in those unimpaired videos ($\overline{MSV}_{blur} = 1.95$ and $\overline{MSV}_{bloc} = 1.09$).



(a) (b)

Figure 5.2: Exp.2s: MSV plots for combinations (bloc;blur): (a) only -blockiness and -blurriness, and (b) blockiness and blurriness.

For sequences impaired with only one artifact, a RM-ANOVA found significant statistical differences between the MSV_{blur} and any pair of only-blurriness (see Table 5.5), and MSV_{bloc} and any pair of only-blockiness (see Table 5.6). These results indicate that participants correctly perceived the different artifact strengths introduced in the videos.

Table 5.5: Exp.2s: Pairwise comparisons between average MSV_{blur} for sequences with only-blurriness (*. Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(0.0;0.4)	(0.0;0.6)	-39.990*	2.631
(0.0;0.4)	(0.0;0.8)	-75.905*	2.125
(0.0;0.6)	(0.0;0.8)	-35.914*	2.641

Table 5.6: Exp.2s: Pairwise comparisons between average MSV_{bloc} for sequences with only-blockiness (*. Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(0.4;0.0)	(0.6;0.0)	-20.390*	2.597
(0.4;0.0)	(0.8;0.0)	-39.552*	2.440
(0.6;0.0)	(0.8;0.0)	-19.162*	1.888

For combinations of blockiness and blurriness (i.e. (0.4;0.4), (0.4;0.6), (0.6;0.4), and (0.6;0.6)), MSV_{bloc} were higher than MSV_{blur} for any pair of combinations (see Figure 5.2 (b)). Again, a RM-ANOVA showed that differences between both MSV_{blur} and MSV_{bloc} for any pair of combinations were statistically significant. An exception were the MSV_{blur} differences for the combination pair (0.4;0.4) and (0.6;0.4) which were not statistically significant (see Table 5.7).

To verify if we can predict the perceived annoyance of videos using the MSVs of blockiness and blurriness, we tested a set of linear and non-linear models on the MSV_{bloc}, MSV_{blur}, and MAV data. The first model was a simple linear model, without any interaction term, given by:

$$PA_{E2,L1} = \alpha \cdot MSV_{bloc} + \beta \cdot MSV_{blur}, \tag{5.3}$$

and the second model a linear model with an intercept term δ, as given by:

$$PA_{E2,L2} = \delta + \alpha \cdot MSV_{bloc} + \beta \cdot MSV_{blur}. \tag{5.4}$$

Table 5.7: Exp.2s: Pairwise comparisons between average MSV_{bloc} and MSV_{blur} for any pair of blurriness and blockiness (*. Significant at 0.05 level.)

Combinations		MSV_{bloc}		MSV_{blur}	
		Diff. Mean	Std. Error	Diff. Mean	Std. Error
(0.4;0.4)	(0.4;0.6)	-12.629*	2.915	-14.133*	3.068
(0.4;0.4)	(0.6;0.4)	-22.638*	2.414	0.248	2.709
(0.4;0.4)	(0.6;0.6)	-34.267*	2.330	-7.590*	3.307
(0.4;0.6)	(0.6;0.4)	-10.010*	2.220	14.381*	2.713
(0.4;0.6)	(0.6;0.6)	-21.638*	2.108	6.543*	2.881
(0.6;0.4)	(0.6;0.6)	-11.629*	1.525	-7.838*	2.906

Table 5.8: Exp.2s: Fitting parameters for linear model without intercept ($PA_{E2,L1}$) (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($>|t|$)	PCC	SCC
α	0.797	0.016	49.220	$< 2e-16*$	0.971	0.958
β	0.721	0.023	30.840	$< 2e-16*$		

Table 5.9: Exp.2s: Fitting parameters for linear model with intercept ($PA_{E2,L2}$). (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($>|t|$)	PCC	SCC
δ	0.386	1.725	0.224	0.824		
α	0.793	0.025	32.302	$< 2e-16*$	0.971	0.958
β	0.716	0.032	22.103	$< 2e-16*$		

For both models, fitting results returned coefficients, α and β, that are statistically significant (Column 5 in Tables 5.8 and 5.9. However, the intercept term (δ) in Equation 5.4 was not found statistically significant. In fact, adding an intercept did not change the values of the correlation coefficients. An ANOVA test showed that differences between $PA_{E2,L1}$ and $PA_{E2,L2}$ models are not statistically significant.

To understand how perceptual artifact strengths interact with one another, we also tested a linear model with interactions, as given by:

$$PA_{E2,L3} = (\alpha \cdot MSV_{bloc} + \beta \cdot MSV_{blur} + \gamma \cdot MSV_{bloc} \cdot MSV_{blur}), \tag{5.5}$$

and, the same model with an intercept coefficient (δ), given by:

$$PA_{E2,L4} = (\delta + \alpha \cdot MSV_{bloc} + \beta \cdot MSV_{blur} + \gamma \cdot MSV_{bloc} \cdot MSV_{blur}). \tag{5.6}$$

Tables 5.10 and 5.11 show results of both fittings. For both models, we can observe that the coefficients (α, β, and γ) are all statistically significant (Column 5 in Tables 5.10 and 5.11). Also, for both models, the correlation coefficients are slightly higher than those for the linear models with no interactions (see Equation 5.3). However, in both models, the interaction term (γ) was negative. These results seem to indicate that there are masking effects among artifacts.

We also tested two Minkowski metrics. The first metric without an intercept term:

$$PA_{E2,M1} = (\alpha \cdot MSV_{blo}^{m} + \beta \cdot MSV_{blu}^{m})^{\frac{1}{m}}, \tag{5.7}$$

Table 5.10: Exp.2s: Fitting parameters for the linear metric with interactions ($PA_{E2,L3}$) (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($>|t|$)	PCC	SCC
α	0.874	0.029	30.059	$< 2e-16*$		
β	0.747	0.024	31.551	$< 2e-16*$	0.975	0.966
γ	-0.004	0.001	-3.105	0.004*		

Table 5.11: Exp.2s: Fitting parameters for the linear metric with interactions and intercept term ($PA_{E2,L4}$). (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($>|t|$)	PCC	SCC
δ	-1.553	1.733	-0.896	0.373		
α	0.899	0.040	22.396	$< 2e-16*$	0.975	0.966
β	0.770	0.035	22.116	$< 2e-16*$		
γ	-0.005	0.001	-3.219	0.002*		

and the second metric with the intercept term:

$$PA_{E2,M2} = (\delta + \alpha \cdot MSV_{blo}^{m} + \beta \cdot MSV_{blu}^{m})^{\frac{1}{m}}, \tag{5.8}$$

where $PA_{E2,M1}$ and $PA_{E2,M2}$ are the predicted annoyance values and m is the Minkowski power obtained from the fit. For both models, we can notice that the coefficients, m, α and β, are statistically significant (Column 5 in Tables 5.12 and 5.13). However, the intercept term (δ) was not found statistically significant. In fact, adding an intercept did not change the values of the correlation coefficients. An ANOVA test showed that differences between $PA_{E2,M1}$ and $PA_{E2,M2}$ models are not statistically significant.

We also predicted annoyance using a SVR model (i.e. $PA_{E2,SVR}$) from MSV_{bloc} and

Table 5.12: Exp.2s: Fitting parameters for Minkowski model ($PA_{E2,M1}$) (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($>|t|$)	PCC	SCC
m	1.341	0.132	10.190	9.99e-16*		
α	0.870	0.029	29.590	$< 2e-16*$	0.975	0.965
β	0.693	0.030	22.820	$< 2e-16*$		

Table 5.13: Exp.2s: Fitting parameters for Minkowski model with intercept ($PA_{E2,M2}$). (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($>\|t\|$)	PCC	SCC
m	1.357	0.137	9.921	3.57e-15*	0.975	0.965
δ	1.692	3.811	0.444	0.658		
α	0.868	0.033	26.543	$< 2e-16*$		
β	0.686	0.032	21.188	$< 2e-16*$		

Table 5.14: Fitting of linear and non-linear models on MAV.

Models	PCC	SCC
$PA_{E2,L1}$	0.971	0.958
$PA_{E2,L2}$	0.971	0.958
$PA_{E2,L3}$	0.975	0.966
$PA_{E2,L4}$	0.975	0.966
$PA_{E2,M1}$	0.975	0.965
$PA_{E2,M2}$	0.975	0.965
$PA_{E2,SVR}$	**0.982**	0.948

MSV_{blur}. Our tests showed that using a radial kernel for the SVR provided the best performance, with PCC and SCC equal to 0.982 and 0.948, respectively. The parameters obtained from SVR are summarized in the third row of Table 5.25. A summary of results obtained for all linear and non-linear models are showed in Table 5.14.

5.4 Experiment 3: Packet-loss, Blockiness and Blurriness

In Exp.3s, we used test sequences with up to three different types of artifacts: packet-loss, blockiness, and blurriness. Again, as shown in Figure 5.3, results showed that MSVs for the combination (0.0;0.0;0.0) (original video) are not equal to zero, indicating that participants perceived impairments in unimpaired videos. Also, in general, participants correctly identified artifacts, giving highest MSVs to the corresponding strongest artifact and smaller MSVs to the other two artifacts.

For combinations with only one artifact, the highest MSVs corresponded to the videos containing only that artifact being verified (see Table 5.15). For example, in a video

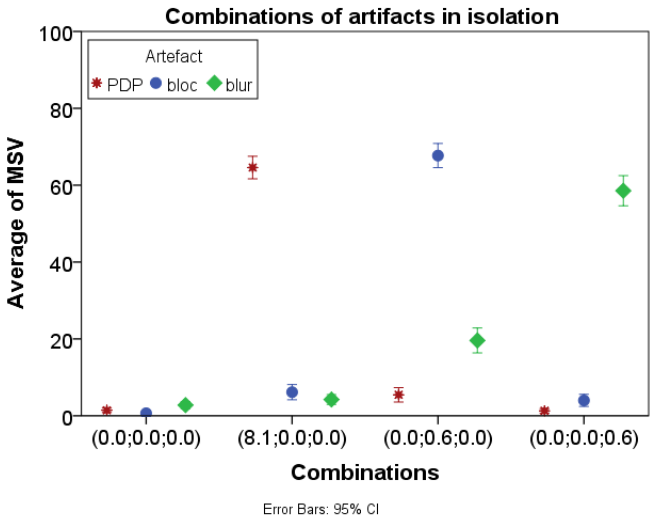


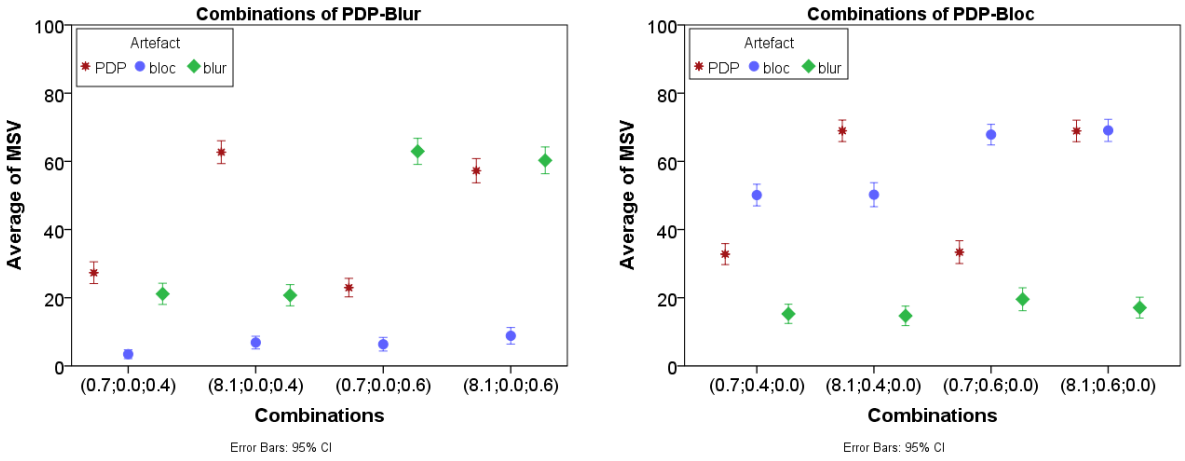
Figure 5.3: Exp.3s: MSV plot combinations (PDP;bloc;blur) for (0.0;0.0;0.0), (8.1;0.0;0.0), (0.0;0.6;0.0), and (0.0;0.0;0.6).

Table 5.15: Exp.3s: Pairwise comparisons between average MSVs for sequences with only -packet-loss, -blockiness, and -blurriness (*. Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(8.1;0.0;0.0)	(0.0;0.0;0.6)	6.029*	2.401
(8.1;0.0;0.0)	(0.0;0.6;0.0)	-3.118	1.801
(0.0;0.6;0.0)	(0.0;0.0;0.6)	9.147*	2.206

impaired with only packet-loss (e.g. $PDP = 8.1\%$) the highest MSV was MSV_{pck}. A RM-ANOVA showed that significant statistical differences in MSV were found for any pair of combinations, except for the combination pair (8.1;0.0;0.0) and (0.0;0.6;0.0). The average MSV was slightly higher for blockiness, followed by packet-loss, and blurriness.

For combinations with two types of artifacts ((*PDP*;*bloc*;0.0), (*PDP*;0.0;*blur*), or (0.0;*bloc*;*blur*)), in most cases, the artifact signal corresponding to the highest signal strength received the highest MSV. Nevertheless, an increase in the strength of a particular artifact signal did not always resulted in a proportional increase in this artifact perceived strength. For example, for (*PDP*;0.0;*blur*) combinations, an increase in the strength of blurriness caused a decrease in the perceived strength of packet-loss artifacts (see Figures 5.4 (a)). A RM-ANOVA found that there are significant statistical MSV differences between MSV_{pck} and all combination pairs of (*PDP*;0.0;*blur*). An exception was found for combination pairs ((0.7;0.0;0.4), (8.1;0.0;0.4)) and ((0.7;0.0;0.6), (8.1;0.0;0.6)), whose MSV_{blur} differences are not statistically significant (see Table 5.16). Notice that, for these two combinations, only the packet-loss strength changed while the blurriness



(a) (b)

Figure 5.4: Exp.3s: MSV plots combinations (*PDP*;*bloc*;*blur*) for (a) (*PDP*;*blur*), and (b) (*PDP*;*bloc*).

Table 5.16: Exp.3s: Pairwise comparisons between average MSVs for (*PDP*;*blur*) sequences (*. Significant at 0.05 level.)

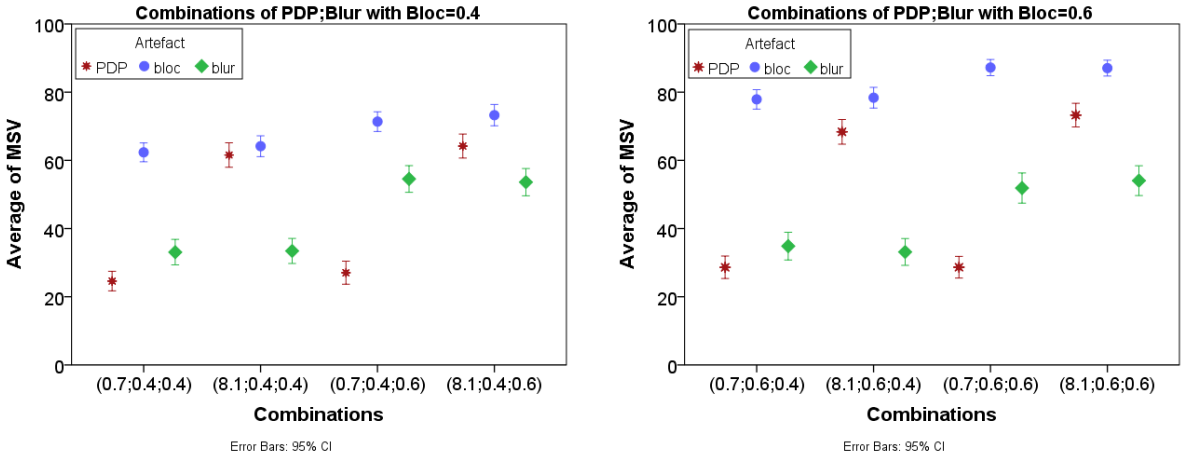
Combinations		MSV_{pck}		MSV_{blur}	
		Diff. Mean	Std. Error	Diff. Mean	Std. Error
(0.7;0.0;0.4)	(8.1;0.0;0.4)	-35.351*	1.645	0.400	1.741
(0.7;0.0;0.4)	(0.7;0.0;0.6)	4.371*	1.636	-41.800*	2.097
(0.7;0.0;0.4)	(8.1;0.0;0.6)	-29.914*	1.753	-39.159*	2.193
(8.1;0.0;0.4)	(0.7;0.0;0.6)	39.722*	1.637	-42.200*	2.137
(8.1;0.0;0.4)	(8.1;0.0;0.6)	5.437*	1.614	-39.559*	2.116
(0.7;0.0;0.6)	(8.1;0.0;0.6)	-34.286*	1.685	2.641	1.843

strength was kept constant. This result suggests that blurriness may be masking the strength of packet-loss artifacts.

The presence of packet-loss in the (*PDP*;*bloc*;0.0) combinations changed the perceived strength of the blockiness artifact (see Figure 5.4 (b)). This indicates that increasing the packet-loss strength in a (*PDP*;*bloc*;0.0) combination can intensify the perceived strength of blockiness. This may be caused by the visual similarity of blockiness and packet-loss artifacts, which are both characterized by the presence of rectangular areas distributed over the video frames. A RM-ANOVA test (see Table 5.17) showed that there are significant statistical differences in MSV_{pck} for all combinations pairs (*PDP*;*bloc*;0.0). The only exceptions are the combination pairs ((0.7;0.4;0.0), (0.7;0.6;0.0)) and ((8.1;0.4;0.0), (8.1;0.6;0.0)). Another RM-ANOVA showed that there are significant

Table 5.17: Exp.3s: Pairwise comparisons between average MSVs for (*PDP*;*bloc*) sequences (*. Significant at 0.05 level.)

Combinations		MSV_{pck}		MSV_{bloc}	
		Diff. Mean	Std. Error	Diff. Mean	Std. Error
(0.7;0.4;0.0)	(8.1;0.4;0.0)	-36.167*	1.652	-0.114	1.885
(0.7;0.4;0.0)	(0.7;0.6;0.0)	-0.576	1.757	-17.718*	1.613
(0.7;0.4;0.0)	(8.1;0.6;0.0)	-36.127*	1.795	-18.959*	1.855
(8.1;0.4;0.0)	(0.7;0.6;0.0)	35.592*	1.779	-17.604*	1.847
(8.1;0.4;0.0)	(8.1;0.6;0.0)	0.041	1.760	-18.845*	1.927
(0.7;0.6;0.0)	(8.1;0.6;0.0)	-35.551*	1.890	-1.241	1.515



(a) (b)

Figure 5.5: Exp.3s: MSV plots combinations (*PDP*;*bloc*;*blur*): (a) (*PDP*;*blur*) with *bloc*=0.4, (b) (*PDP*;*blur*) with *bloc*=0.6.

statistical differences in MSV_{bloc} values for all combination pairs, except for combination pairs ((0.7;0.4;0.0), (8.1;0.4;0.0)) and ((0.7;0.6;0.0), (8.1;0.6;0.0)).

For combinations that correspond to videos with the three types of artifact signals, the average MSV_{bloc} was higher than the average MSV_{pck} and MSV_{blur}, respectively. Figures 5.5 (a) and (b) show plots of combinations with different values of packet-loss, blockiness, and blurriness strengths. A RM-ANOVA showed that there are significant statistical differences between MSVs for most combinations of (*PDP*;*bloc*;*blur*).

The combination pairs ((0.7;0.4;0.4), (0.7;0.4;0.6)) and ((8.1;0.4;0.4), (8.1;0.4;0.6)) were not found to having statistically significant differences in MSV_{pck}. Although only the strength of blurriness varied in both combination pairs, we verified that MSV_{bloc} also increased as MSV_{blur} increased. This result suggests that the blockiness is affected by

Table 5.18: Exp. 3: Pairwise comparisons between average MSVs for sequences with blockiness=0.4 and changing packet-loss and blurriness strengths (*. Significant at 0.05 level.)

Combinations		MSV_{pck}			MSV_{bloc}			MSV_{blur}		
		Diff. Mean	Std. Error		Diff. Mean	Std. Error		Diff. Mean	Std. Error	
(0.7;0.4;0.4)	(8.1;0.4;0.4)	-36.976*	1.768		-1.788	1.501		-0.351	1.712	
(0.7;0.4;0.4)	(0.7;0.4;0.6)	-2.445	1.781		-9.000*	1.437		-21.478*	2.049	
(0.7;0.4;0.4)	(8.1;0.4;0.6)	-39.608*	1.833		-10.902*	1.664		-20.531*	2.161	
(8.1;0.4;0.4)	(0.7;0.4;0.6)	34.531*	1.969		-7.212*	1.491		-21.117*	2.028	
(8.1;0.4;0.4)	(8.1;0.4;0.6)	-2.633	1.699		-9.114*	1.767		-20.180*	2.171	
(0.7;0.4;0.6)	(8.1;0.4;0.6)	-37.163*	1.916		-1.902	1.570		0.947	2.198	

an increase of the blurriness strength. For MSV_{bloc} and MSV_{blur}, the differences of both combination pairs ((0.7;0.4;0.4),(8.1;0.4;0.4)) and ((0.7;0.4;0.6), (8.1;0.4;0.6)) are not statistically significant. Although only packet-loss strength changes, MSV variations were higher for MSV_{bloc} than for MSV_{blur} (see Table 5.18 columns 5 and 7). These results support the assumption that the packet-loss artifact can intensify the perception of the blockiness than blurriness.

When comparing MSVs for sequences with *bloc*=0.6 and different *PDP* and *blur* values (*PDP*;0.6;*blur*), a RM-ANOVA showed that, for most combination pairs, differences are statistically significant. For MSV_{pck}, only the difference for the combination pairs ((0.7;0.6;0.4), (0.7;0.6;0.6)) was not statistically significant. Differences in both MSV_{bloc} and MSV_{blur} were statistically significant for all combination pairs, with the exception of combination pairs ((0.7;0.6;0.4), (8.1;0.6;0.4)) and ((0.7;0.6;0.6), (8.1;0.6;0.6)), as shown in columns 5 and 7 of Table 5.19.

We tested a set of linear and non-linear models, fitting them on MSV_{pck}, MSV_{bloc}, MSV_{blur}, and the MAV data. The first linear model was a simple linear model, without any interaction term:

$$PA_{E3,L1} = \alpha \cdot MSV_{pck} + \beta \cdot MSV_{bloc} + \gamma \cdot MSV_{blur}. \tag{5.9}$$

Next, we adapted Equation 5.10 to include an intercept coefficient (δ):

$$PA_{E3,L2} = \delta + \alpha \cdot MSV_{pck} + \beta \cdot MSV_{bloc} + \gamma \cdot MSV_{blur}. \tag{5.10}$$

Table 5.19: Exp. 3: Pairwise comparisons between average MSVs for sequences with *bloc*=0.6 and changing packet-loss and blurriness strengths (*. Significant at 0.05 level.)

Combinations		MSV_{pck}		MSV_{bloc}		MSV_{blur}	
		Diff. Mean	Std. Error	Diff. Mean	Std. Error	Diff. Mean	Std. Error
(0.7;0.6;0.4)	(8.1;0.6;0.4)	-39.710*	1.942	-0.482	1.485	1.714	2.147
(0.7;0.6;0.4)	(0.7;0.6;0.6)	-0.020	2.085	-9.327*	1.249	-17.029*	2.312
(0.7;0.6;0.4)	(8.1;0.6;0.6)	-44.616*	1.990	-9.151*	1.310	-19.208*	2.339
(8.1;0.6;0.4)	(0.7;0.6;0.6)	39.690*	1.921	-8.845*	1.306	-18.743*	2.327
(8.1;0.6;0.4)	(8.1;0.6;0.6)	-4.906*	1.605	-8.669*	1.323	-20.922*	2.119
(0.7;0.6;0.6)	(8.1;0.6;0.6)	-44.596*	1.854	0.176	1.031	-2.180	2.430

Tables 5.20 and 5.21 show the fitting results for both models. Notice that all coefficients (i.e. δ, α, β, and γ) are statistically significant and the correlation coefficients are greater than 0.90.

Since we are also interested in understanding if the perceptual strengths interact with one another, we tested a linear model with interactions without an intercept term, as

Table 5.20: Fitting parameters for linear model without intercept ($PA_{E3,L1}$) (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($>\|t\|$)	PCC	SCC
α	0.340	0.022	18.330	$< 2e-16*$		
β	0.470	0.020	23.210	$< 2e-16*$	0.937	0.936
γ	0.413	0.026	16.04	$< 2e-16*$		

Table 5.21: Fitting parameters for linear model with intercept ($PA_{E3,L2}$). (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($>\|t\|$)	PCC	SCC
δ	3.846	1.870	2.057	0.042*		
α	0.370	0.026	14.313	$< 2e-16*$	0.937	0.937
β	0.456	0.021	21.448	$< 2e-16*$		
γ	0.371	0.033	11.326	$< 2e-16*$		

Table 5.22: Fitting parameters for the linear metric with interactions $PA_{L3,E3}$ (* Significant at 0.05 level).

Coefficient	Estimate	Std. Error	t-value	Pr ($>\|t\|$)	PCC	SCC
α	5.476e-01	3.572e-02	15.327	$< 2e-16*$		
β	5.470e-01	4.535e-02	12.062	$< 2e-16*$		
γ	4.432e-01	3.530e-02	12.558	$< 2e-16*$		
ρ_1	-2.918e-03	1.054e-03	-2.768	0.006*	0.956	0.947
ρ_2	-3.414e-03	1.321e-03	-2.585	0.011*		
ρ_3	-1.855e-04	1.277e-03	-0.145	0.885		
ρ_4	1.908e-05	2.834e-05	0.673	0.502		

given:

$$\begin{aligned} PA_{E3,L3} = \alpha \cdot MSV_{pck} + \beta \cdot MSV_{bloc} + \gamma \cdot MSV_{blur} + \rho_1 \cdot MSV_{pck}MSV_{bloc} \\ +\rho_2 \cdot MSV_{pck}MSV_{blur} + \rho_3 \cdot MSV_{bloc}MSV_{blur} + \rho_4 \cdot MSV_{pck}MSV_{bloc}MSV_{blur}. \end{aligned} \tag{5.11}$$

We also adapted the Equation 5.11 to include an intercept coefficient (δ):

$$\begin{aligned} PA_{E3,L4} = \delta + \alpha \cdot MSV_{pck} + \beta \cdot MSV_{bloc} + \gamma \cdot MSV_{blur} + \rho_1 \cdot MSV_{pck}MSV_{bloc} \\ +\rho_2 \cdot MSV_{pck}MSV_{blur} + \rho_3 \cdot MSV_{bloc}MSV_{blur} + \rho_4 \cdot MSV_{pck}MSV_{bloc}MSV_{blur}. \end{aligned} \tag{5.12}$$

Tables 5.22 and 5.23 show the fitting results for both models. Notice that most first, second, and third order coefficients are statistically significant. The exceptions are ρ_3 and ρ_4 in $PA_{E3,L3}$ model in Table 5.22, which correspond to the interaction of (*bloc*;*blur*) and (*PDP*;*bloc*;*blur*), respectively. Notice also that most second order coefficients were negative, what may indicate masking effects, i.e. when two artifacts are present, one of them may attenuate the strength of the others artifact(s). The interaction coefficient with highest magnitude corresponded to (*PDP*;*blur*). This suggests that packet-loss artifacts affect how blurriness artifacts are perceived. Again, the correlation coefficient values are all greater than 0.950.

Next, we tested the weighted Minkowski metric, which included weights for each individual artifact strength, as given by the following equation:

$$PA_{E3,M1} = (\alpha \cdot MSV_{pck}^m + \beta \cdot MSV_{bloc}^m + \gamma \cdot MSV_{blur}^m)^{\frac{1}{m}}, \tag{5.13}$$

where α, β, and γ are the weights for MSV_{pck}, MSV_{bloc}, and MSV_{blur}, respectively, and m

Table 5.23: Fitting parameters for the linear metric with interactions and an intercept term $PA_{L3,E4}$ (* Significant at 0.05 level).

Coefficient	Estimate	Std. Error	t-value	Pr ($>|t|$)	PCC	SCC
δ	-1.857e+01	2.768e+00	-6.710	5.22e-10*		
α	8.516e-01	5.488e-02	15.516	$< 2e-16*$		
β	8.411e-01	5.888e-02	14.286	$< 2e-16*$		
γ	7.670e-01	5.713e-02	13.424	$< 2e-16*$	0.965	0.957
ρ_1	-7.729e-03	1.161e-03	-6.654	6.93e-10*		
ρ_2	-8.740e-03	1.393e-03	-6.274	4.66e-09*		
ρ_3	-5.488e-03	1.360e-03	-4.036	9.17e-05*		
ρ_4	1.062e-04	2.778e-05	3.821	0.000*		

Table 5.24: Fitting parameters for the Minkowski model $PA_{L3,M1}$ (* Significant at 0.05 level).

Coefficient	Estimate	Std. Error	t-value	Pr ($>|t|$)	PCC	SCC
m	1.993	0.143	13.960	$< 2e-16*$		
α	0.387	0.023	17.130	$< 2e-16*$	0.969	0.963
β	0.565	0.021	27.760	$< 2e-16*$		
γ	0.321	0.029	11.280	$< 2e-16*$		

is the Minkowski power. Table 5.24 shows the fitting results. Notice that all coefficients are statistically significant (Columns 5 in Table 5.24). Blockiness is the artifact with the highest impact on MAV, followed by packet-loss and blurriness. Again, the correlation coefficient values are all greater than 0.950. Figure 5.6 shows a plot of the observed MAV versus $PA_{E3,M1}$, using the parameters obtained from the fit.

Finally, we use SVR to predict annoyance (i.e. $PA_{E3,SVR}$) from MSV_{pck}, MSV_{blo} and MSV_{blu}. Table 5.25 summarizes the SVR results over all Experiments, with columns 2-5 showing the estimated parameters and columns 6-7 showing the PCC and SCC values for the fit.

To compare models in terms of the trade-off between complexity and accuracy, we use the Akaike Information Criterion (AIC). For the different models, different degrees of accuracy were achieved, yet in some cases at the expense of increased complexity. For example, models with interaction terms have more parameters to be fit than models without these terms. As a consequence, although more accurate, they may be more prone to

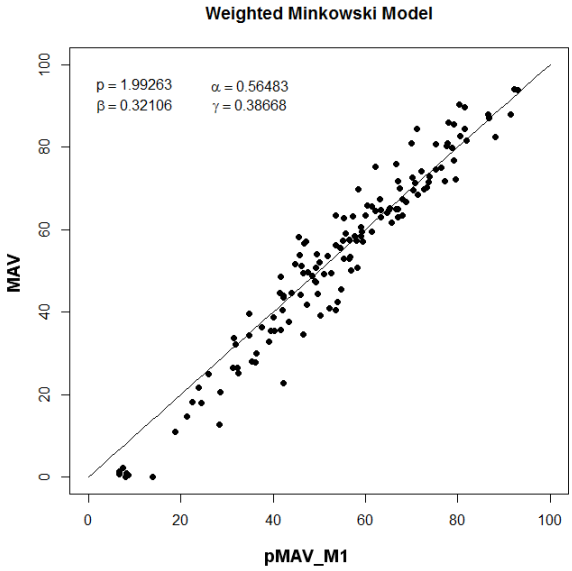


Figure 5.6: Exp 3: Observed MAV versus predicted MAV using the weighted Minkowski metric ($PA_{E3,M1}$) for the data set containing all test videos.

Table 5.25: Fitting parameters for SVR model by Experiments.

Experiment	K	C	γ	ϵ	PCC	SCC
1s	radial	64	1	0.0	0.953	0.927
2s	radial	8	0.5000	0.0	**0.982**	0.948
3s	radial	4	0.3333	0.1	0.963	**0.957**

overfitting. Table 5.26 summarizes the AIC values computed for all models, except for the SVR model. A model with lower AIC is preferred, what means the $PA_{E2,M1}$ (Minkowski model tested in Exp.2s) has the compromise between performance and complexity.

To compare the other models with the SVR model, we used (again) a 10-fold cross-validation setting to obtain comparable measurements. Results of this test are depicted in Table 5.27. Unfortunately, we were not able to perform this test with the Minkowski models. Notice that $PA_{E2,L3}$ outperforms all models in terms of AIC. However its performance in terms of correlation is similar to the linear model with interactions and with an intercept term.

Correlation coefficients for all models are summarized in Table 5.27. Experimental results showed that there are interactions among artifact signals. Therefore, while designing quality models, it is important to take this into consideration to avoid underestimating or overestimating quality. Nevertheless, results also showed that, although annoyance cannot be predicted from a single artifact measurement, it is frequently much better to

Table 5.26: Akaike Information Criterion (AIC) for the linear and Minkowski models. A lower value indicates a better trade-off between model complexity and accuracy.

Exp.1s			Exp.2s			Exp.3s		
Model	df	AIC	Model	df	AIC	Model	df	AIC
$PA_{E1,L1}$	2	627.473	$PA_{E2,L1}$	3	517.883	$PA_{E3,L1}$	4	984.615
$PA_{E1,L2}$	3	622.126	$PA_{E2,L2}$	4	519.831	$PA_{E3,L2}$	5	982.327
			$PA_{E2,L3}$	4	510.451	$PA_{E3,L3}$	8	949.302
			$PA_{E2,L4}$	5	511.609	$PA_{E3,L4}$	9	910.218
			$PA_{E2,M1}$	4	**509.032**	$PA_{E3,M1}$	5	907.273
			$PA_{E2,M2}$	5	510.720			

Table 5.27: Average correlation across the 10-fold cross-validation runs between model predictions and MAVs

Model	PCC	SCC	Model	PCC	SCC	Model	PCC	SCC
$PA_{E1,L1}$	0.955	0.935	$PA_{E2,L1}$	0.968	0.912	$PA_{E3,L1}$	0.938	0.917
$PA_{E1,L2}$	0.955	0.935	$PA_{E2,L2}$	0.968	0.912	$PA_{E3,L2}$	0.938	0.918
			$PA_{E2,L3}$	0.975	0.929	$PA_{E3,L3}$	0.951	0.929
			$PA_{E2,L4}$	0.975	0.926	$PA_{E3,L4}$	0.975	0.926
$PA_{E1,SVR}$	0.953	0.927	$PA_{E2,SVR}$	**0.982**	0.948	$PA_{E3,SVR}$	0.963	**0.957**

use a subset of the most significant artifacts to predict annoyance.

5.5 Annoyance Models based on Artifact Metrics

In this section, we investigated if we can use a combination of artifact metrics to predict annoyance. In other words, we used the same linear and non-linear models described in the previous section to combine the outputs of NR artifact metrics (see Table 2.2).

5.5.1 Experiment 1

We adapted the linear models described in Equations 5.1 and 5.2 to predict MAV using NR artifact metrics. Table 5.28 shows PCC, SCC, and AIC values obtained from each NR metric. Although the test sequences of Exp.1s contain only packet-loss artifacts, we can notice that $Bloc_F$ has the best performance in terms of AIC and correlation coefficients.

Table 5.28: Exp.1s: PCC, SCC, and AIC values obtained using a set of artifact metrics to predict annoyance, with the linear models in Eqs. 5.1 and 5.2.

Metrics	$PA_{E1,L1}$			$PA_{E1,L2}$		
	PCC	SCC	(df, AIC)	PCC	SCC	(df, AIC)
$Pack_B$	-0.147	0.109	(2, 913.905)	0.147	-0.109	(3, 838.250)
$Pack_R$	0.278	0.297	(2, 885.229)	0.278	0.300	(3, 832.923)
$Blur_M$	-0.243	-0.352	(2, 854.479)	0.243	0.341	(3, 834.711)
$Blur_N$	0.226	0.310	(2, 833.986)	0.226	0.311	(3, 835.472)
$Blur_C$	-0.190	-0.336	(2, 850.925)	0.190	-0.336	(3, 836.878)
$Bloc_W$	-0.066	0.109	(2, 880.007)	0.066	-0.109	(3, 839.845)
$Bloc_F$	**0.362**	**0.308**	**(2, 831.899)**	**0.362**	**0.308**	**(3, 827.435)**

Table 5.29: Exp.1s: PCC and SCC obtained using $Bloc_F$ and $Pack_R$ metrics to predict annoyance, with the $PA_{E1,SVM}$ model (SVR algorithm).

Metrics	Cost	γ	ϵ	PCC	SCC
$Pack_R$	8	1	0.9	0.353	0.251
$Bloc_F$	512	1	0.9	0.486	0.435

Among packet-loss metrics, $Pack_R$ has the best performance in terms of AIC and correlation coefficients. So, we tested these two metrics using the SVR model. Table 5.29 shows the parameters used, as well as, the correlation coefficients obtained for each metric. Notice that $Bloc_F$ has the best performance in terms of correlation coefficients. On the other hand, the correlation values obtained with all metrics are very low.

5.5.2 Experiment 2

First, we tested the set of NR artifact metrics on test sequences containing only-blockiness artifacts. Table 5.30 shows the correlation coefficients and the AIC values obtained for each metric. We can notice that $Bloc_F$ has the best performance in terms of correlation and AIC values, for both $PA_{E2,L1}$ and $PA_{E2,L2}$ models, with PCC values higher 0.80.

Next, we tested the set of NR artifact metrics on test sequences containing only-blurriness artifacts. Table 5.31 shows the results obtained for each metric. For these sequences, $Blur_C$ had the best performance in terms of AIC and correlation, for both

Table 5.30: Exp.2s: PCC, SCC, and AIC values for the linear models considering all NR artifact metrics for only-blockiness sequences.

Metrics	$PA_{E2,L1}$			$PA_{E2,L2}$		
	PCC	SCC	AIC	PCC	SCC	AIC
$Pack_B$	0.080	0.142	(2, 213.246)	0.080	0.143	(3, 183.081)
$Pack_R$	0.511	0.497	(2, 218.749)	0.511	0.497	(3, 176.853)
$Blur_M$	0.283	0.236	(2, 180.628)	0.283	0.236	(3, 181.457)
$Blur_N$	-0.316	-0.244	(2, 186.059)	0.316	0.244	(3, 181.004)
$Blur_C$	0.211	0.276	(2, 181.540)	0.211	0.276	(3, 182.256)
$Bloc_W$	-0.228	-0.261	(2, 209.539)	0.228	0.261	(3, 182.094)
$Bloc_F$	**0.806**	**0.865**	**(2, 170.926)**	**0.806**	**0.865**	**(3, 161.157)**

Table 5.31: Exp.2s: PCC, SCC, and AIC values for the linear models considering all NR artifact metrics for only-blurriness sequences.

Metrics	$PA_{E2,L1}$			$PA_{E2,L2}$		
	PCC	SCC	AIC	PCC	SCC	AIC
$Pack_B$	-0.288	-0.315	(2, 224.506)	0.288	0.315	(3, 204.350)
$Pack_R$	-0.347	-0.391	(2, 224.942)	0.347	0.457	(3, 203.483)
$Blur_M$	0.545	0.477	(2, 197.090)	0.545	0.477	(3, 198.762)
$Blur_N$	-0.540	-0.449	(2, 212.637)	0.540	0.449	(3, 198.920)
$Blur_C$	**0.630**	**0.536**	**(2, 196.769)**	**0.630**	**0.536**	**(3, 195.564)**
$Bloc_W$	-0.449	-0.465	(2, 218.482)	0.449	0.465	(3, 201.434)
$Bloc_F$	0.494	0.439	(2, 203.436)	0.494	0.439	(3, 200.299)

$PA_{E2,L1}$ and $PA_{E2,L2}$ models. Notice that all blurriness metrics seem to perform well when used in video sequences impaired with only blurriness.

Then, we used the metrics $Bloc_F$ (blockiness) and $Blur_C$ (blurriness) using the linear models in Equations 5.3 to 5.6 on test sequences containing combinations of blockiness and blurriness artifacts. These two metrics were selected because of their good performance for only-blockiness and only-blurriness sequences. Table 5.32 shows the PCC, SCC, and AIC values obtained for this combination of metrics ($Bloc_F, Blur_C$).

Notice that the model $PA_{E2,L4}$ provided the best PCC value, with all coefficients being statistically significant (Columns 5 in Table 5.33, except for β coefficient (referring

Table 5.32: Exp.2s: PCC, SCC, and AIC values for the linear models considering $Bloc_F$ and $Blur_C$.

Models	PCC	SCC	AIC
$PA_{E2,L1}$	0.765	0.830	**(3, 204.132)**
$PA_{E2,L2}$	0.765	0.825	(4, 206.006)
$PA_{E2,L3}$	0.765	0.823	(4, 206.120)
$PA_{E2,L4}$	**0.796**	**0.827**	(5, 204.582)

Table 5.33: Fitting parameters for the linear metric ($PA_{E2,L4}$) with interactions, an intercept term with $Bloc_F$ and $Blur_C$ as parameters, for test sequences with only combination of blockiness and blurriness (* Significant at 0.05 level).

Coefficient	Estimate	Std. Error	t-value	Pr ($>\|t\|$)
δ	-195.50	108.70	-1.798	0.0848*
α	198.40	78.80	2.518	0.0189*
β	681.50	403.20	1.690	0.1039
ρ_1	-508.20	287.60	-1.767	0.0900*

to $Blur_C$ metric). Although, ρ_1 coefficient is statistically significant, it has a negative value, indicating that masking effects may be present.

Next, we used the combination of these metrics ($Bloc_F, Blur_C$) for predicting MAV considering all test sequences of Exp.2s, for the linear models and the SVR algorithm. Among the linear models, $PA_{E2,L4}$ had the best performance in terms of correlation coefficients (PCC = 0.827 and SCC = 0.862) and AIC value (df = 5, AIC = 653.098). For the SVR, we used a radial kernel, with Cost = 16, $\gamma = 0.5$, and $\epsilon = 0.4$, with PCC = 0.852 and SCC = 0.758.

In summary, we noticed that $Bloc_F$ and $Blur_C$ seem to be the best predictors for blockiness and blurriness, respectively. When tested in combination to predict MAV considering all test sequences of Exp.2s, results showed that the $PA_{E2,L4}$ model had the best performance in terms of correlation coefficients and AIC values. Also, all coefficients were statistically significant (see Table 5.34). Although ρ_1 is statistically significant, its negative value might indicate that masking effects among the artifacts are present.

Table 5.34: Fitting parameters for the linear metric ($PA_{E2,L4}$) considering all test sequences of Exp.2s, with $Bloc_F$ and $Blur_C$ as parameters (* Significant at 0.05 level).

Coefficient	Estimate	Std. Error	t-value	Pr ($>|t|$)
δ	-438.66	78.17	-5.612	3.40e-07*
α	395.44	67.96	5.819	1.47e-07*
β	1301.11	275.63	4.720	1.11e-05*
ρ_1	-1040.26	240.09	-4.333	4.63e-05*

5.5.3 Experiment 3

Taking into account the previously results obtained, we tested the combination ($Pack_R$, $Blur_C$, $Bloc_F$) metrics in the sequences of Exp.3s. We chose these metrics because they presented the best performance results in Exp.1s and Exp.2s. Table 5.35 shows the correlation coefficients and AIC values obtained for each model. Results show that the combination function that achieves the best performances is $PA_{E3,M1}$ (Eq. 5.13). We also used SVR model to predict annoyance from the combination of the $Pack_R$, $Blur_C$, and $Bloc_F$ metrics. For this test, we ran the SVR using a radial kernel, with a gathered Cost = 4, $\gamma = 0.33$, and $\epsilon = 0.20$, obtaining PCC = 0.820 and SCC = 0.804.

Table 5.35: Exp.3s: PCC, SCC, and AIC values for all model investigated.

Models	$Pack_R, Bloc_F, Blur_C$		
	PCC	SCC	AIC
$PA_{E3,L1}$	0.763	0.819	(4, 1154.147)
$PA_{E3,L2}$	0.778	**0.848**	(5, 1146.791)
$PA_{E3,L3}$	0.764	0.819	(8, 1148.461)
$PA_{E3,L4}$	0.795	0.844	(9, 1144.648)
$PA_{E3,M1}$	**0.806**	0.841	**(5, 1130.710)**

In summary, the model with the best performance, in terms of correlation and AIC values, was $PA_{E3,M1}$ model. Correlation coefficients obtained for the Minkowski model were higher than those obtained for the linear models. The Minkowski power found ($m = 0.164$) was considerably lower than the values found by other authors [10, 18], but very close the value found earlier [34, 79]. This may indicate that the model based on artifact metrics is, in fact, more sensitive to small changes than the model based on perceptual strengths. Finally, the SVR-based approach has the best performance among

all models. This indicates that linear and non-linear models were not able to capture the complex non-linear processes that may be part of this combination model [79, 89].

5.6 Discussion

We presented the methodology, statistical analysis, and conclusions of the strength task performed in the three psychophysical experiments. The goal was to model the overall annoyance using the perceptual strengths of blockiness, blurriness, and packet-loss artifacts. The models allow us to understand how these artifacts combine and interact to produce overall annoyance. The results showed that, when artifact signals were presented alone at a high strength, participants were able to identify them correctly. At low strengths, on the other hand, other artifacts were reported. Annoyance increased with both the number of artifacts and their strength.

Annoyance models were obtained by combining the artifact perceptual strengths (MSV) using a weighted Minkowski model, a support vector regression (SVR) model, and a linear model on the experimental data. Performing an ANOVA test, we found that all types of artifact signal strengths had a significant effect on MAV. The ANOVA test also indicated that there are interactions among some of the artifact perceptual strengths. We also tested a non-linear model using SVR. This provided greater correlation coefficients than using other models. In summary, annoyance can be modeled as a multidimensional function of the individual artifact signal measurements [10, 14, 45, 46].

These results indicate that a NR quality model based on artifact measurements is indeed a valid approach, but it needs to include a minimal set of relevant artifacts. Also, although annoyance cannot be predicted using only one individual artifact signal measurement, it is not necessary to use all possible artifacts and it would suffice to use the most significant artifacts (perceptually). For example, blockiness seems to have the highest effect on the predicted MAV. Finally, results show that there are interactions among artifact signals. Therefore, while designing quality models, it is important to take this into consideration to avoid underestimating or overestimating quality.

Annoyance models were obtained by combining the artifact metrics to predict annoyance are tested and, they have showed a lower performance than annoyance models by combining the artifact perceptual strengths. In general, the $Bloc_F$ artifact metric seems to have a higher performance among other artifact for packet-loss and blockiness artifacts, whilst $Blur_C$ seems to have a higher performance for blurriness artifacts. Among all tested models, the SVR-based approach has the best performance among all of them with a correlation coefficient higher than 0.820.

Chapter 6

Visual Attention

In this chapter, we examined the viewing behavior during both quality assessment tasks and free-viewing of videos impaired with multiple artifacts. More specifically, we aimed at detecting differences in 1) fixation duration and 2) spatial gaze allocation for videos containing combinations of blockiness, blurriness, and packet-loss. We reported the outcomes of an eye-tracking study during which observers were asked to freely look at pristine videos and score the annoyance of a set of impaired versions of those videos. The resulting eye-tracking data are converted into saliency information (i.e. saliency maps averaged across all participants for each video and under each viewing condition) and analyzed changes in gaze locations due to both task and artifact annoyance.

6.1 Introduction

Recent studies show that the assessment of video quality is closely tied to gaze deployment [57]. When observing a scene, the human eye typically scans the video neglecting areas carrying little information, while focusing on visually important regions [58]. Wang *et al.* [59] showed that, within the first 2,000ms of observation, gaze patterns target main objects in a scene. Later, the gaze is redirected to other salient, yet not visually important, areas. This result suggests that visual coding should be focused, at first, into the main objects of the scene. Nevertheless, the presence of artifacts may disrupt these natural gaze patterns, causing viewer's annoyance and, consequently, lower quality judgments [60]. Therefore, saliency information should be incorporated into video quality metrics.

Several researches in the area of visual quality have tried to incorporate gaze pattern information into the design of visual quality metrics [61], mostly using the assumption that visual distortions appearing in less salient areas might be less visible and, therefore, less annoying [62, 93]. However, while some researchers report that the incorporation of gaze pattern information increases the performance of quality metrics, others report no

or very little improvement [63]. One possible reason for such disagreement is that, still, the role played by visual attention in quality evaluation is unclear. Although it has been shown that, for images, artifacts in visually important regions are far more annoying than those in the background [64], it is still not clear if artifacts can create saliency (and therefore, attract gaze) on their own. And if so, it is unclear which type of artifacts can create saliency and at what perceptual strength. If artifacts can disrupt gaze patterns by creating saliency, this should be taken into account in the design of quality metrics that make use of saliency or gaze pattern information. Unfortunately, the existing knowledge in this direction is scattered.

Ninassi *et al.* [65] studied viewing behavior during both free-viewing and quality assessment of impaired images. They found two results: (1) the quality task has a significant effect on the fixation duration, which increased on unimpaired images during a quality scoring task and, (2) the type of impairment degrading the image causes modifications in gaze patterns. Le Meur *et al.* [67] examined viewing behavior during both quality assessment and free-viewing tasks. Differently from images, they found that the average fixation duration is almost the same for both tasks; whereas saliency does not change significantly when videos are impaired (coding artifacts). Redi *et al.* [68] investigated to what extent the presence of packet-loss artifacts influences viewing behavior. However, contrary to Le Meur *et al.* [67], they showed that saliency can significantly change in free-viewing and quality assessment tasks and these changes are related to content and impairment strength. Similarly, Mantel *et al.* [69] found a positive correlation between coding artifacts annoyance and fixation dispersion.

From these results, it seems that, for both images and videos, some artifacts (e.g. packet-loss) may be able to divert gaze and viewing behavior from their natural paths. But, it is yet unclear when and how this happens. It is important to point out that most studies have focused on analyzing the impact that artifacts in isolation have on gaze patterns [68–71]. In real-life situations, it is very likely that different artifacts are co-present in a video. For example, packet-loss may occur in the transmission of a severely compressed video, creating perceptual degradations that are very different from the single artifacts in isolation. To the best of our knowledge, there is no study that explores the impact of combinations of artifacts on gaze patterns and viewing behavior.

So, we examined the viewing behavior during both quality assessment and free-viewing of videos impaired with multiple artifacts of videos from Exp.3a. More specifically, we analyzed differences in: 1) fixation duration and 2) spatial gaze allocation for videos containing combinations of blockiness, blurriness, and packet-loss. We reported the outcomes of an eye-tracking study during which observers were asked to freely view at pristine videos and score the annoyance of a set of impaired versions of those videos. The resulting eye-

tracking data are converted into saliency information (i.e. saliency maps averaged across all participants for each video and under each viewing condition) and analyzed to detect any changes in gaze locations due to both task and artifact annoyance.

6.2 Experimental Results

Taking into account the annoyance of multiple artifacts, Figure 6.1 represents the average MAV per video content, averaged across all 19 combinations. We can notice that the MAV are slightly different, across videos: the videos *Into Tree* (49.19 ± 19.18), *Romeo and Juliet* (50.27 ± 21.71) and *Cactus* (46.90 ± 20.62) obtained a relatively smaller MAV than *Park Joy* (56.38 ± 21.33), *Park Run* (56.62 ± 24.73), *Basketball* (57.14 ± 20.92) and *Barbecue* (53.45 ± 22.31). These findings are in line with previous works available in literature, which observed that video content influences MAVs [68, 94].

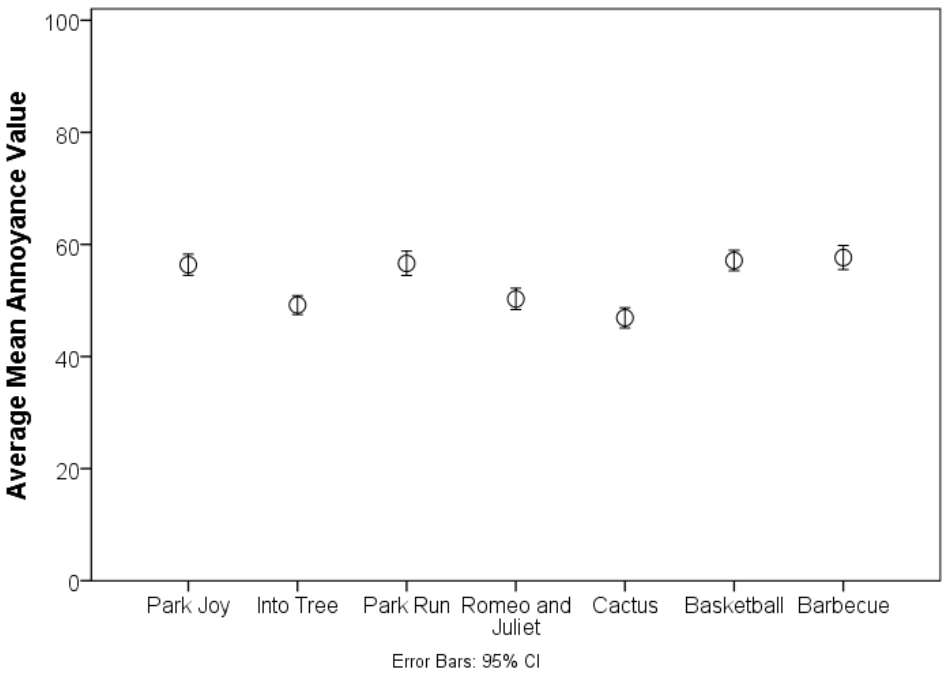


Figure 6.1: Average MAV computed over all the distorted versions of each video.

Nevertheless, in this study the impact of video content on MAV seems to be smaller than what was found by Redi *et al.* [68] using the same videos and the same packet-loss artifacts. A possible cause for this difference is that we use two additional artifacts (blockiness and blurriness) instead of only packet-loss, as in Redi's work. The annoyance of blocky and blurry videos may depend less on the temporal characteristics of the video.

Figure 6.2 shows the distribution of MAV across the several combinations of artifacts used in the test (SC_{PV}, SC_{G1}, SC_{G2} and SC_{G3}). The average MAV for SC_{G1} (39.67 ±

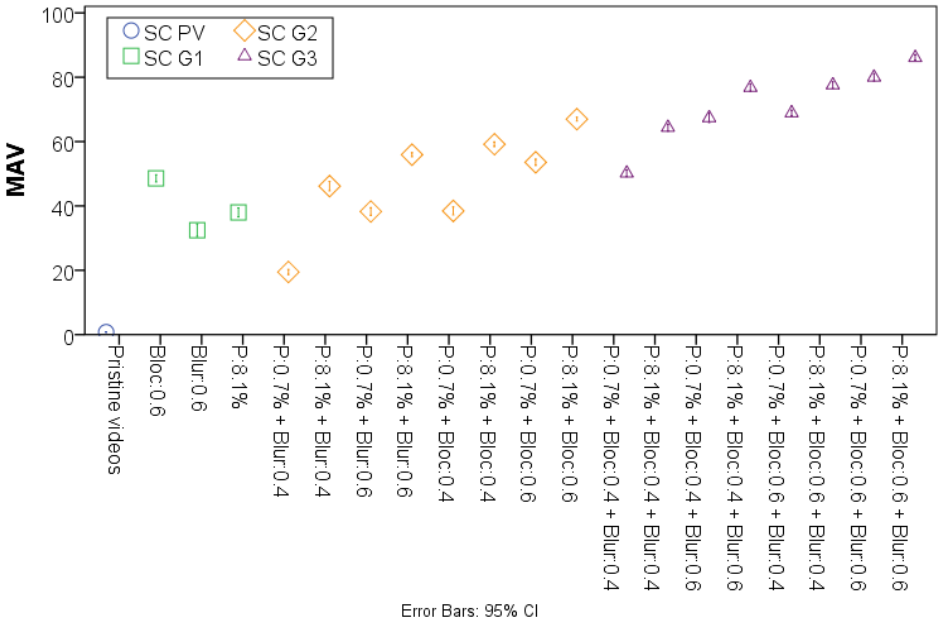


Figure 6.2: Average MAV over all videos for all combinations of artifacts (see Table 3.3).

12.26) is lower than for SC_{G2} (47.26 $\pm$ 15.58) and SC_{G3} (71.40 $\pm$ 12.92). An ANOVA showed that these differences are statistically significant (F = 1475.98, df = 2, $p = .000$), except for combination (0.7;0.0;0.4) with a lower MAV than the combination (0.0;0.0;0.6), which had the smallest MAV of SC_{G1}.

Maybe, in this case, the presence of packet-loss artifacts is masking the blurriness. Another exception is combination (0.0;0.6;0.0) that is more annoying than half of the combinations in SC_{G2}, which are mostly combinations of packet-loss and blurriness artifacts. This suggests that the blockiness artifact, when in isolation or in the presence of other artifacts, is more annoying than the other two artifacts.

6.2.1 Fixation duration

In order to analyze the viewing behavior, we studied the fixation duration recorded during free-viewing and quality assessment tasks. Figure 6.3 shows the average fixation duration per video content for both tasks. A RM-ANOVA using the video groups as the independent variables and the fixation duration as the dependent variable did not find a significant difference between the average fixation duration for free-viewing and quality assessment tasks (F(5.809, 116.175) = 2.113, $p = 0.059$).

These results suggest that the average fixations duration is similar when free-viewing and quality assessment tasks are considered, which contradicts the finding of Redi *et al.* [68], who showed significant differences in the average duration fixations between free-viewing and quality assessment tasks. Our result is instead in line with what found by Le

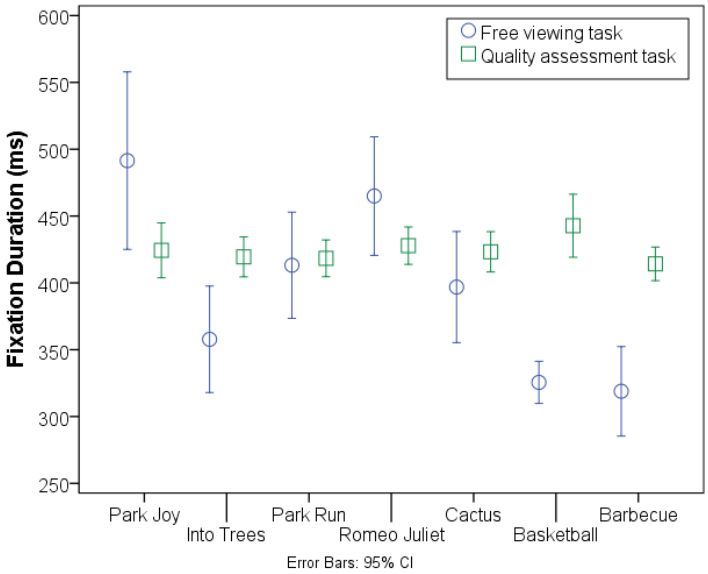


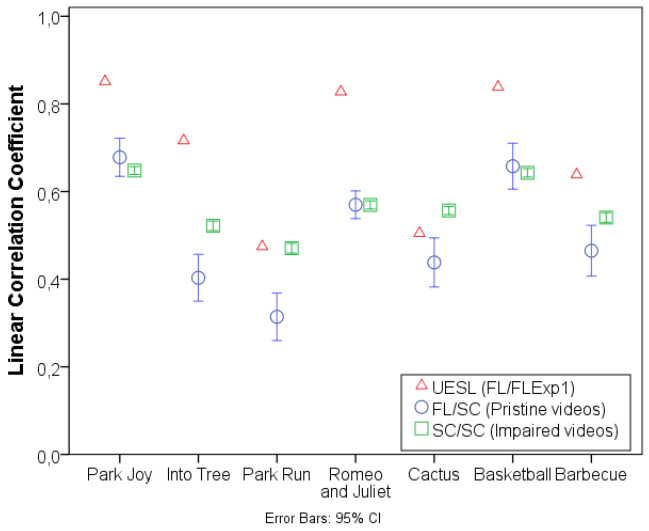
Figure 6.3: Average fixation duration for free-viewing (blue circles) and quality assessment (green squares) tasks.

Meur *et al.* [70]. As in our experiment, videos in Le Meur *et al.* included blocky artifacts, whereas videos in Redi *et al.* were impaired by packet-loss. We may hypothesize then that packet-loss artifacts cause an increase in the duration of fixation, but since in our experiment packet-loss artifacts were overlapping with blocking artifacts (which seem not to impact on fixation duration), their effect may be reduced.

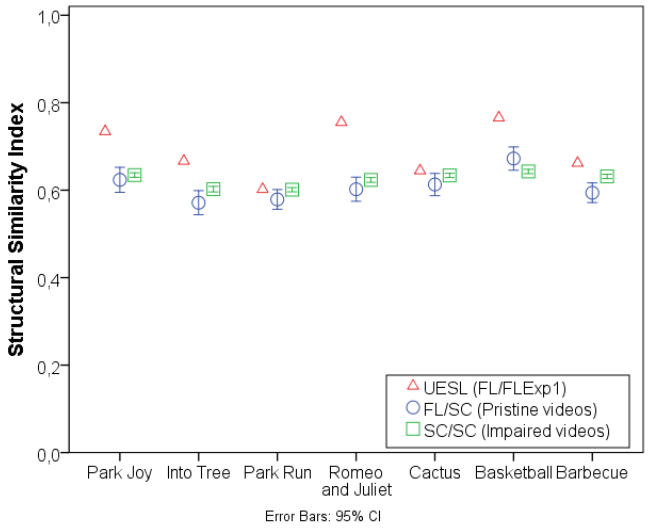
6.2.2 Similarities among saliency maps

To take a closer look at the viewing behavior for sequences with combinations of blockiness, blurriness, and packet-loss artifacts, we looked into the similarities between saliency maps for free-viewing and quality assessment tasks. Figure 6.4 shows (a) LCC and (b) SSIM between saliency maps computed for the same video under different tasks: quality assessment (SC_{PV}) and free-viewing (FV_{PV}). The UESL values are also included to represent inter-observer variability. Notice from this graph that the similarity between quality assessment and free-viewing maps is systematically lower than the UESL, showing that task does have an impact on viewing behavior.

To check whether the presence of artifacts altered the viewing behavior, we calculated the similarity between saliency maps obtained for pristine and impaired videos, during a quality assessment task. We considered all combinations used in SC_{G1}, SC_{G2}, and SC_{G3} (133 impaired videos used in quality assessment tasks) as a single video group (SC_{SC}). Since the saliency distributions are gathered for the same task (quality assessment), we expected the similarity measures for SC_{SC} to be close to the UESL values. From Figure



(a)



(b)

Figure 6.4: (a) LCC and (b) SSIM Similarity measures computed between maps obtained from pristine videos during free-viewing and quality assessment tasks.

6.4 we can notice that the saliency of pristine and impaired videos (quality assessment task) is different. Similarity measures are lower than UESL values, showing that the presence of artifacts (in this case, combined) alters saliency maps. These results are in agreement with what was found by Redi *et al.* [68].

Next, we verified if this attention change is influenced by the different combinations of artifacts. Figure 6.5 (a) shows the similarity (in terms of LCC) between the FV_{PV} maps and the SC_{PV} maps. Additionally, it shows the similarity between the maps corresponding to free-viewing of videos and the maps for the scoring of the corresponding video impaired with all different combinations of artifacts (thus, belonging to $FV_{SC_{G1}}, FV_{SC_{G2}}$ and $FV_{SC_{G3}}$). It can be seen from this figure that, at least quantitatively, the change in

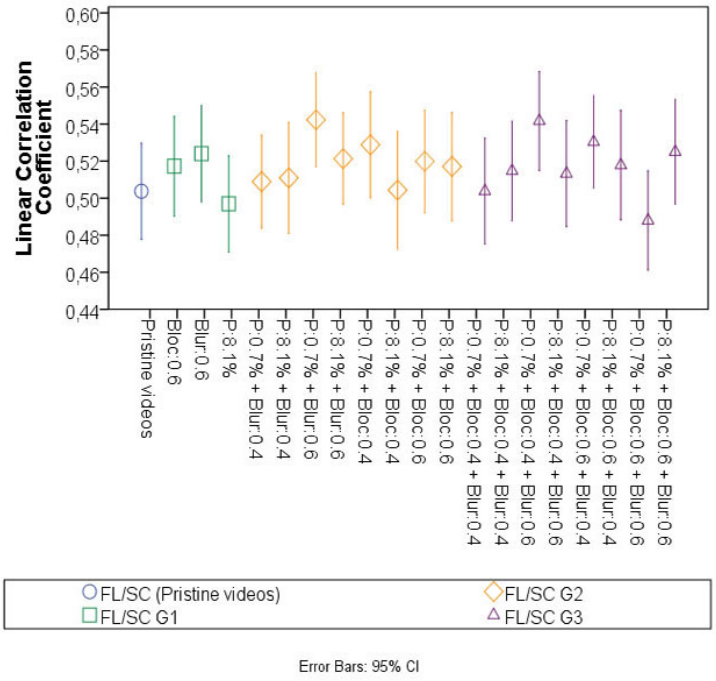
saliency is independent of the number or the type of artifacts. An ANOVA revealed that the similarities among saliency maps for FV_{SC} and $FV_{SC_{G1}}, FV_{SC_{G2}}$, and $FV_{SC_{G3}}$ were not statistically significant (F = 0.971, df = 19, $p = 0.493$).

Similarly, we compared the saliency maps for FV_{SC} and $SC_{SC_{G1}}, SC_{SC_{G2}}$, $SC_{SC_{G3}}$. As shown in Figure 6.5 (b), the specific combination also does not seem to play a role in the saliency changes. Compared to the impact of task (see the blue mark on the left side of Figure 6.5 (b) corresponding to the comparison of the saliency maps of pristine videos for free-viewing and quality assessment tasks), the impact of specific artifact combinations seems negligible. An ANOVA revealed indeed that the similarity for FV_{SC} is significantly lower than that of $SC_{SC_{G1}}, SC_{SC_{G2}}$ and $SC_{SC_{G3}}$ (F = 3.155, df = 19, $p = 0.000$).

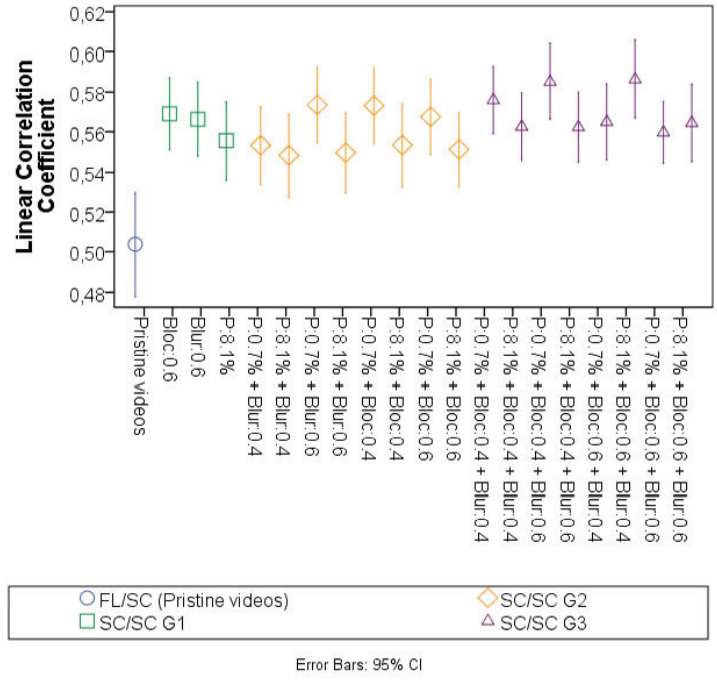
Although the specific artifact combinations do not seem to have an impact on gaze locations, there may still be a relationship between the perceived quality of the video and saliency distribution. To check this hypothesis, we measured the similarity of saliency maps of pristine and impaired videos in quality assessment task. Figure 6.6 shows how $LCC(SC_{PV}, SC_{SC})$ varies depending on the MAV of the impaired videos. We considered 3 categories of MAV: $MAV < 30$, $30 < MAV < 60$ and $MAV > 60$. An ANOVA revealed that the LCC between these maps is significantly different among categories of MAV (F = 10.483, df = 2, $p = 0.000$). Notice that the similarity among saliency maps obtained from scoring pristine and impaired videos increases with the annoyance of the artifacts.

6.3 Discussion

We studied the effect combinations of artifacts have on viewing behavior. With this goal, we tracked eye movements of 21 participants while they were watching videos impaired with combinations of blockiness, blurriness, and packet-loss. Then, we analyzed the viewing behavior of our participants in terms of fixation durations and spatial gaze allocation. Our results indicated that the presence of impairments had no impact on the duration of fixations. Nevertheless, analyzing saliency maps we were able to detect changes in gaze deployment. In particular, we measured the similarity of saliency maps corresponding to the same video captured for different tasks (free-viewing or quality assessment), types of impairments (different combinations of packet-loss, blockiness and blurriness), or a combination of the two. Our results show that differences in viewing behavior exist due to a change in task. Also, the presence of impairments in the video impacts the saliency distribution. We did not find an effect of a specific type of artifact combination on saliency changes.



(a)



(b)

Figure 6.5: Similarity among saliency maps computed LCC for (a) pristine videos during free-viewing and quality assessment tasks and (b) pristine and impaired videos during quality assessment tasks.

Interestingly, the similarity measure of the saliency maps increased with the increase of the artifact annoyance. This is a counter intuitive result, as one would expect more

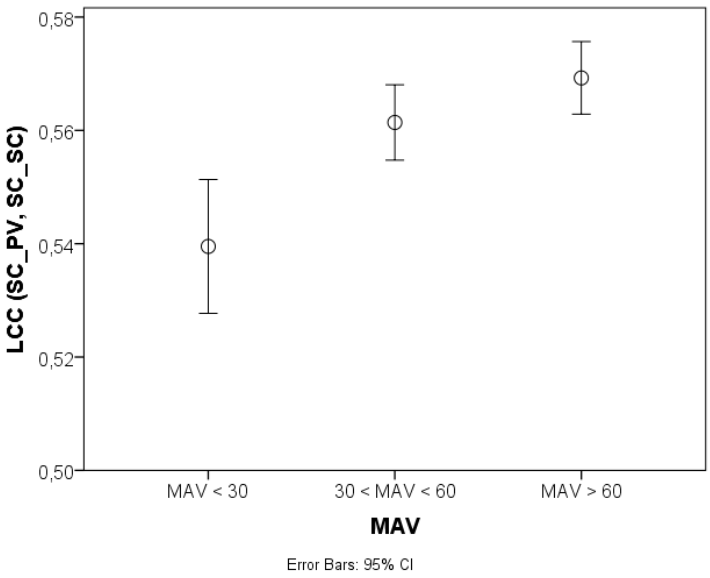


Figure 6.6: LCC Similarity among saliency maps obtained scoring pristine and impaired videos, for different categories of MAV.

annoying artifacts to be visually stronger and, thus, create saliency on their own. A possible explanation for this result is that, whereas for combinations with low MAV the impact of localized packet-loss artifacts is more evident, for more annoying combinations, this localized effect may have been masked by the presence of other artifacts. So, the source of annoyance may become indistinctively diffuse across the whole video area. For this reason, further analysis is needed to link the change in saliency to physical properties of the video. Two main points need to be addressed in the future. First, we need to establish whether the change in saliency detected by our similarity metrics relates to the divergence of fixations outside the core region of interest of the video or it comes from a convergence of fixations within it. This may indicate if artifacts create saliency on their own. Second, it is necessary to further understand the link between saliency changes and artifact type, annoyance, and location.

Chapter 7

Proposed Video Quality NR Metric

In this chapter, we presented the proposed NR video quality assessment method, which is inspired in the experimental results analyzed in the earlier chapters. More specifically, it combines artifact and visual attention *features* to produce an estimate of the video quality.

7.1 Introduction

Video quality metrics (VQM) aim to predict the perceived video quality automatically (i.e. without human intervention), using different approaches [95]. For example, Mean Square Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) measure the accuracy of the video signal without modeling any aspect of the HVS [36]. Although, they are quite simple and widely used for video quality, usually these data metrics cannot give a quality measure that correlates well with the perceived quality [35, 36].

NR video quality metrics are usually designed using either vision model or engineering approaches [96,97]. Vision modeling approaches (or perceptual oriented metrics) take into account the most important HVS characteristics. Engineering approaches (the so-called top-down approaches) are based on the extraction structural image characteristics (e.g. contours) or artifacts (e.g. blockiness or blurriness which are introduced by a particular compression method or transmission link), measuring the strength of these *features* to estimate the overall quality [95].

As mentioned earlier, the design of NR metrics is a challenging task. So, understanding that the perceived quality of a video can be affected by a variety of artifacts and that the strengths of these artifacts contribute to the overall annoyance is an important contribution. It is worth mentioning that measuring individual artifact strengths can be much faster and accurate than trying to directly estimate the overall annoyance [11].

7.2 Proposed Method

In this work, we use the *features* of three of the most relevant artifacts found in digital videos (blockiness, blurriness, and packet-loss) to blindly predict video quality [34]. More specifically, based on the experimental results presented in Chapters 4 to 6, we tested a couple of models using *features* extracted with artifact metrics. The *features* were extracted from the spatial and frequency domain. Since the Discrete Cosine Transform (DCT) has an important role in several video applications (e.g. JPEG, MPEG, and H.261 codecs), frequency *features* were extracted from the DCT domain.

This approach has some similarities with some approaches currently available in the literature. For example, Marichal *et al.* [98] proposed a new technique to determine blurriness by exploring the available DCT information (based on histograms of non-zero DCT occurrences) in MPEG and JPEG compressed video or images, achieving a low computational cost. Caviedes *et al.* [99] proposed a no-reference metric that measures sharpness using statistical measures of the frequency distribution. This metric proved to be very precise to measure the relative sharpness of multiple versions of the same scene. The algorithm uses a combination of the spatial and frequency measures. Ichigaya *et al.* [100] proposed a no-reference method that estimates the quality of MPEG-2 videos analyzing the DCT coefficients.

Bhattacharyya *et al.* [101] proposed an algorithm for efficient detection of corrupted macroblocks (MB), which calculates the average luminance values of the DCT coefficients of each MB and uses a threshold to detect the presence of edges. Their algorithm provided a considerable improvement in performance. One limitation is the fact that the authors tested their algorithm using only one video sequence and, therefore, the algorithm may be sequence dependent.

In this work, we proposed a NR metric for estimating the overall video quality, which uses a combination of artifact-based *features*, i.e. *features* that characterize the presence and strength of artifacts. More specifically, the method considers *features* like DCT information (i.e. DC and AC coefficients), cross-correlation of sub-sampled images, average absolute differences between block image pixels, intensity variation between neighbouring pixels, and visual attention. A non-linear SVR regression model is used to combine all *features* to obtain an overall quality estimate. Figure 7.1 shows a block diagram of proposed NR video quality assessment model. In the next sections, we described the process for extracting the artifact-based features.

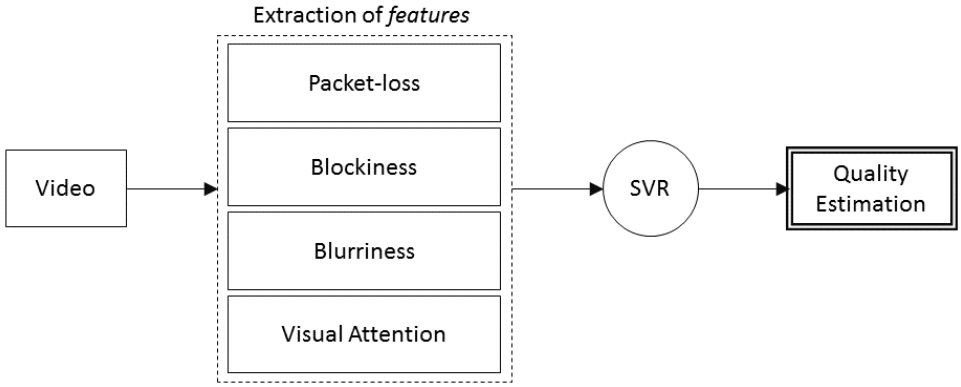


Figure 7.1: Block diagram of a multidimensional no-reference video quality metric, based on a combination of artifact-based *features*.

7.2.1 Packet-loss Features

In this section, we described the algorithm designed to extract *features* that characterize packet-loss perceived distortions. We designed these *features* inspired on the work of Bhattacharyya *et al.* [101], where both DC and AC DCT coefficients can be used to detect strong edges in a frame. Their algorithm computes the average energy of regions composed of neighboring blocks, which are known to be strongly correlated, and compares the DC coefficient of each block with its immediate spatial neighbors.

In this work, we extracted the packet-loss based *features* using two stages (see Figure 7.2): detection and extraction.

The detection stage is divided into the 3 following steps:

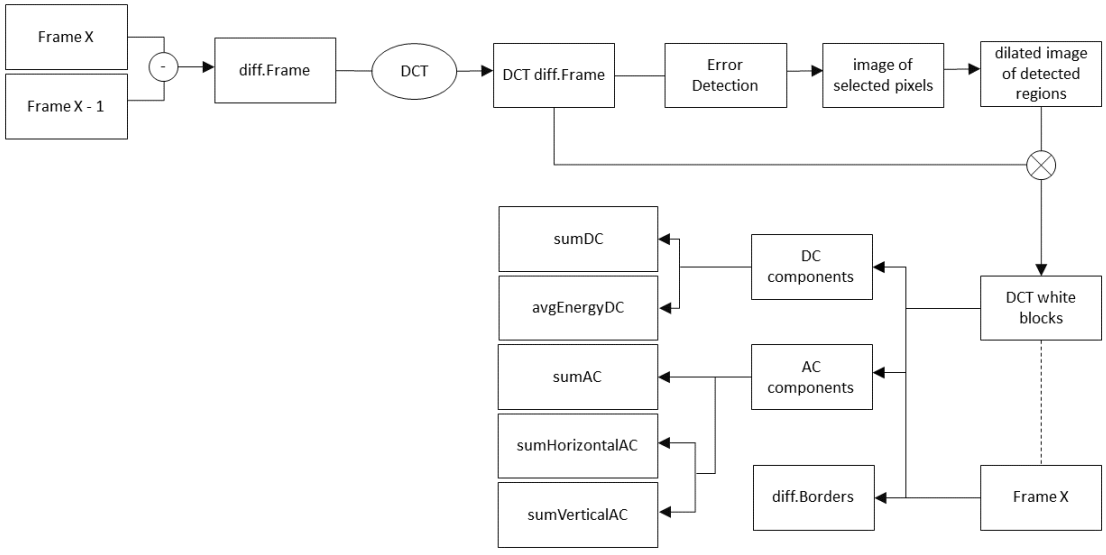


Figure 7.2: Block diagram of complete procedure for packet-loss feature extraction.

1. The difference between the current and previous frames is calculated. The resulting image is named *diff.Frame*;

2. The *diff.Frame* image is split in 8×8 blocks and the DCT of each block is computed. The resulting image in named *DCT diff.Frame*;

3. The average energy of each block is computed. Then, the DC energy of each block is compared with the energy of the block below it. The AC coefficients are also used to extract the edge information. If the module of the DC energy difference and of the sum of the first five AC coefficients are greater than an threshold (say $T = 50$), then a strong edge was found.

Therefore, as a result, 3 images are created:

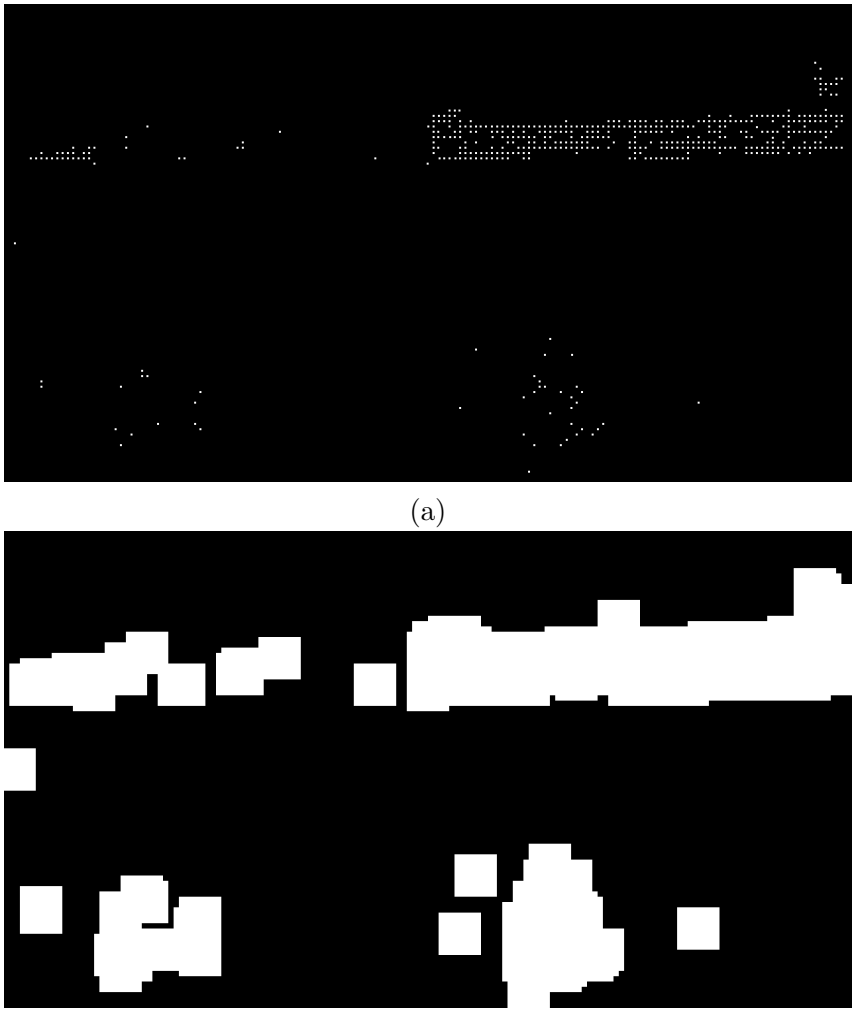
1. The *image of selected pixels* - The white points in this image (see Figure 7.3 (a)) correspond to strong edges, i.e. pixels for which the module of the DC energy difference and the sum of the first five AC coefficients are greater than 50.

2. The *dilated image of detected regions* - This image (see Figure 7.3 (b)) is generated by expanding each white point in the *image of selected pixels* into a 64×64 block.

3. The *DCT white blocks* - This image is generated by multiplying the *DCT diff.Frame* image by the *dilated image of detected regions*.

Next, the extraction procedure is performed in the *DCT white blocks* image. This image is divided in 8×8 blocks and only the DC coefficients that correspond to strong edges (i.e. where DC components $\neq 0$) are considered. So, 6 *features* are extracted from the DCT white blocks image: *avgEnergyDC*, *sumDC*, *sumAC*, *sumVerAC*, *sumHorAC*, and *diffBorders*. These *features* are computed as follows:

- *avgEnergyDC* - Computed by taking the average of all DC coefficients;
- *sumDC* - Computed by summing all the DC coefficients. Notice that there is only one DC component per block (represented in green in Figure 7.4).
- *sumAC* - Computed by summing the absolute values of all first five AC coefficients (yellow squares in Figure 7.4).
- *sumVerAC* - Computed by summing the absolute values of all AC vertical coefficients (red line in Figure 7.4).
- *sumHorAC* - Computed by summing the absolute values of all AC horizontal coefficients (blue line in Figure 7.4).



(a)

(b)

Figure 7.3: Frame 81 of *Intro Tree* video (Exp.1a): images generated from DCT coefficient based error detection process.

- *diffBorders* - Computed by summing the differences between the pixel intensities at the top and bottom of the borders of each block:

$$DB = \sum_{i=1}^{col} |T_1(i) - B_1(i)| + |T_2(i) - B_2(i)| \tag{7.1}$$

  where T_1 and T_2 corresponds to the top lines, B_1 and B_2 to the bottom lines, and *col* is the number of columns of the block (assuming sizes of 8, 16, and 32).

Notice that only detected blocks (DC coefficients correspond to strong edges) are computed in Equation 7.1, all other pixels are assumed zero. Both detection and extraction stages are executed for blocks of 8×8, 16×16, and 32×32 pixels. Therefore, in total we have 18 *features* (6 *features* from 8×8 blocks + 6 *features* from 16×16 blocks + 6 *features* from 32×32 blocks).

Figure 7.4: 8×8 block structure used to compute the DC and AC coefficients, as well as, horizontal and vertical *features*.

7.2.2 Blockiness Features

Results presented in Chapter 5 suggest that Farias' algorithm [14] is able to detect blockiness artifacts with a fairly good performance. In their algorithm, the frame is divided into 8×8 blocks and down-sampled into two separate parts, corresponding to the vertical and horizontal directions.

Figures 7.5 (a) and (b) depict the vertical and horizontal sampling structures for the horizontal and vertical directions, respectively, considering a 24×24 frame area. The dark symbols inside the grid correspond to pixels in the resulting sampled sub-images, with different symbols corresponding to different sub-images. The sub-images sampled only in the horizontal direction (s_h) are given by:

$$s_h(m,n) = \{Y(i,j) : m = i, n = j \mod 8\}, \tag{7.2}$$

while the sub-images sampled only in the vertical direction (s_v) is given by:

$$s_h(m,n) = \{Y(i,j) : m = i \mod 8, n = j\}. \tag{7.3}$$

where (i,j) are the horizontal and vertical co-ordinates. A total of 6 sampled sub-images are obtained after the downsampling process, i.e. 3 sub-images from the vertical downsampling (s_{v_7}, s_{v_0}, s_{v_1}) and 3 sub-images from the horizontal downsampling (s_{h_7}, s_{h_0}, s_{h_1}). Notice that the subscript values attached on s_v and s_h labels represent the pixel position inside the grid in the resulting sampled sub-images.

In the vertical direction, the inter-block correlation is performed by computing the correlation between the sub-images s_{v_7} and s_{v_0} by:

$$PV_{inter} = \max_{i,j} \left\{ C_{s_{v_7}, s_{v_0}}(i,j) \right\}. \tag{7.4}$$

The intra-block correlation is given by

$$PV_{intra} = \max_{i,j} \left\{ C_{s_{v_0}, s_{v_1}}(i,j) \right\}. \tag{7.5}$$

In these equations, $C_{s_{v_7}, s_{v_0}}$ and $C_{s_{v_0}, s_{v_1}}$ give the cross-correlation between two images, which is calculated using the following equation:

$$C_{m,n}(i,j) = F^{-1}\left(\frac{F^*(s_m(i,j)) \cdot F(s_n(i,j))}{|F^*(s_m(i,j))F(s_n(i,j))|} \right), \tag{7.6}$$

where F and F^{-1} denote the forward and inverse 2-D discrete Fourier transform, respectively, and $*$ denotes the complex conjugate. The magnitude of the highest peak is a measure of the correlation between s_m and s_n. Before the maximum value is taken, the array elements are filtered using a Hamming window, what forces the elements around the borders to a constant value. The horizontal correlations, PH_{inter} and PH_{intra}, are obtained in a similar way.

Since Farias *et al.* target only MPEG-2 compression artifacts, their algorithm only takes into account 8×8 blocks. But, given that modern codecs (e.g. H.264 and H.265) also include block sizes of 16×16 and 32×32 pixels, in this work we adapted their algorithm to also include these block sizes. In our metric, these above procedures are also run for 16×16 and 32×32 blocks.

To generate the *features*, we use the following equations for each block size:

$$P_{inter} = PV_{inter} + PH_{inter}$$

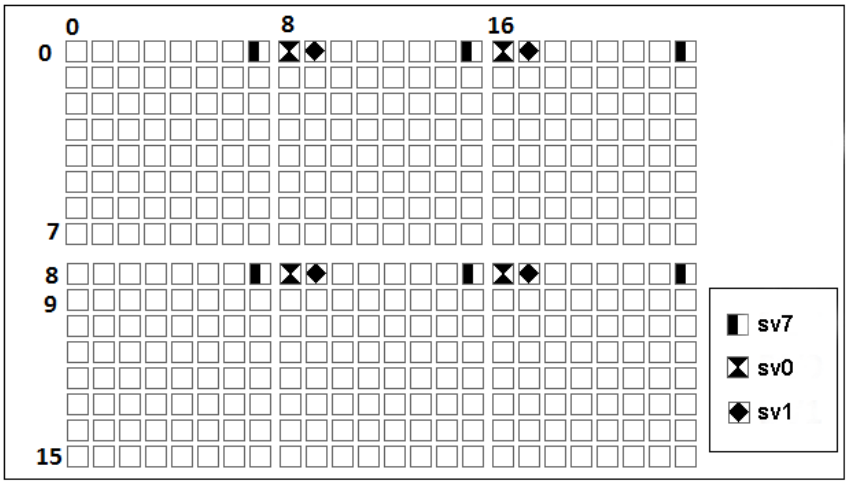
and

$$P_{intra} = PV_{intra} + PH_{intra}.$$

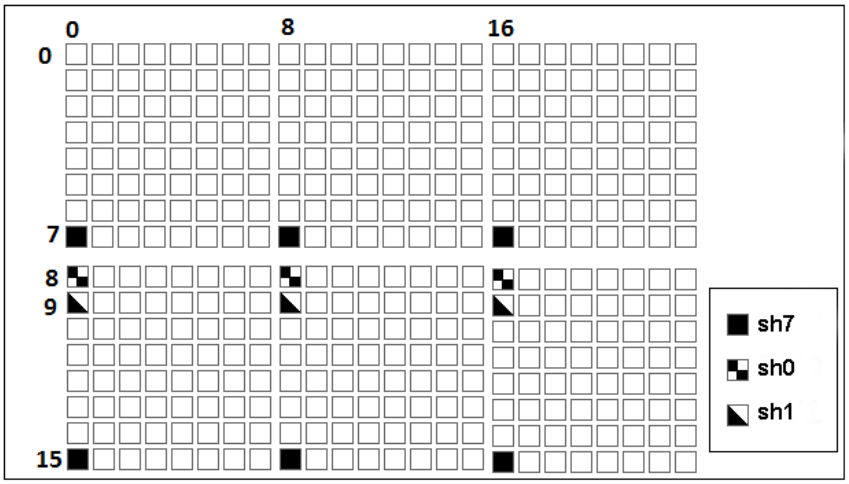
Therefore, we generate a total of 6 *features*: 2 *features* from 8×8 blocks, 2 *features* from 16×16 blocks, and 2 *features* from 32×32 blocks).

7.2.3 Blurriness Features

This feature extraction procedure is based on the blurriness metric proposed by Crete *et al.* [55]. The authors observed that humans have difficulties perceiving differences between



(a)



(b)

Figure 7.5: Sample of frame downsampling structure for 8×8 block size: (a) vertical and (b) horizontal.

a blurred image and the same re-blurred image. Therefore, they estimate the blur strength by comparing the input image with a very blurry version of it (obtained with a strong low-pass filter). The algorithm analyzes the intensity variation of neighboring pixels, taking into account only the pixels that have changed after the blurring step. They also analyzed the pixels containing noise information, which can be located in edges, textured areas, or homogeneous areas [55]. An overview of Crete's algorithm is shown in Figure 7.6.

Similarly to Crete's algorithm, we analyzed only the luminance component of the video frame, Y. A strong filtering is performed horizontally and vertically, generating the following blurred images:

$$B_v = h_v * Y$$

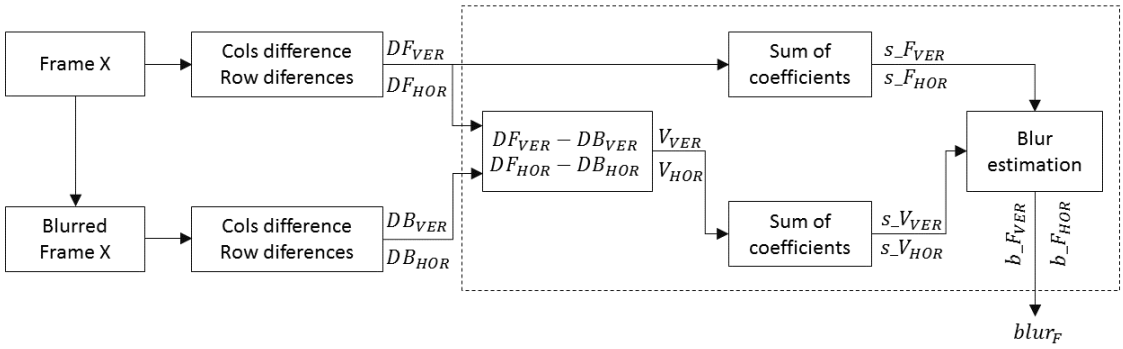


Figure 7.6: Block-diagram of the algorithm to estimation of blur annoyance.

and

$$B_h = h_h * Y,$$

where $h_v = h_h^T = \frac{1}{9}[1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1]$. Then, the variation of the neighboring pixels, after blurring, is evaluated using the following equations:

$$PV_H(k) = \sum_{m=1}^{row-1} \sum_{n=1}^{col-1} \max(0, |Y(m,n) - Y(m,n-1)| - |Y_{blur_h}(m,n) - Y_{blur_h}(m,n-1)|) \tag{7.7}$$

and

$$PV_V(k) = \sum_{m=1}^{row-1} \sum_{n=1}^{col-1} \max(0, |Y(m,n) - Y(m-1,n)| - |Y_{blur_v}(m,n) - Y_{blur_v}(m-1,n)|) \tag{7.8}$$

where row is number of rows, col is number of columns, and $k \leq NF$ is the frame index. If the variation is high, the input image is sharp, whilst if the variation is small, the input image is blurred.

To compare input and blurred images, we compute the sum of the absolute differences of these images:

$$PS_H(k) = \sum_{m=1}^{row-1} \sum_{n=1}^{col-1} |Y(m,n) - Y(m,n-1)| \tag{7.9}$$

and

$$PS_V(k) = \sum_{m=1}^{row-1} \sum_{n=1}^{col-1} |Y(m,n) - Y(m-1,n)|. \tag{7.10}$$

Although in Crete's algorithm the final blur value is simply given by the most annoying blur value in the vertical and horizontal directions, in our metric we use PV_H, PV_V, PS_H, and PS_V as our *blurriness features*.

7.2.4 Visual Attention Features

In a recent work, Zhang *et al.* [102] investigated the capabilities and limitations of the state-of-the-art saliency models for image quality. They found that, although current saliency models yield statistically significant gains in the performance of image quality metrics (IQM), performance gains vary among individual combinations of saliency models and IQMs. Also, the effectiveness of a saliency-based IQM depends on the type of distortion. Zhang *et al.* [103] proposed a FR IQM that uses visual saliency (VS) as a feature, which includes computing the local quality map of the distorted image. Also, when pooling the quality score, VS is used as a weighting function.

In our work, we used a well-known algorithm for visual attention, named the graph-based visual saliency (GBVS) to generate saliency maps for each video frame [104]. GBVS extracts image features, such as, intensity, color, and orientation (similar to Itti *et al.* [105]). All grid locations of each feature map are used to built a fully-connected graph, where weights between two nodes are proportional to the similarity of the feature values and their spatial distance. The dissimilarity between two positions (i, j) and (p, q) in the feature map (FM), with respective feature values $FM(i, j)$ and $FM(p, q)$, is defined as:

$$d\left((i,j);(p,q)\right) = \left|\log \frac{FM(i,j)}{FM(p,q)}\right|. \tag{7.11}$$

The directed edge from node (i, j) to node (p, q) is assigned a weight proportional to their dissimilarity and their distance on lattice FM:

$$w\left((i,j);(p,q)\right) = d\left((i,j);(p,q)\right) \cdot R(i-p, j-q), \tag{7.12}$$

where, $R(a, b) = \exp\left(-\frac{a^2+b^2}{2\sigma^2}\right)$.

The weights of the outbound edges of each node in the resulting graphs are normalized to 1. An equivalence relationship between nodes and states are defined, as well as between edge weights and transition probabilities. Their equilibrium distribution is adopted as the activation and saliency maps. In the equilibrium distribution, large values are assigned to nodes that are highly dissimilar to surrounding nodes. The activation maps are normalized to emphasize details and, then, combined into a single overall map. Once the saliency maps information generated, they are incorporated into our proposed metric by

$$SalMap = \sum_{m=1}^{row} \sum_{n=1}^{col} SM(m, n) \tag{7.13}$$

where $SalMap$ is a similarity measurement [103] and $SM(m, n)$ is the saliency map pixel. Zhang *et al.* [103] proposed a full-reference image quality assessment method that using

the visual saliency as a feature when computing the local quality map of the distorted image, and as a weighting function to reflect the importance of a local region.

7.2.5 Feature Combination

Once we have extracted the packet-loss, blockiness, blurriness, and visual attention *features*, a total of 29 *features* (18 for packet-loss, 6 for blockiness, 4 for blurriness, and 1 for visual attention) are obtained. Then, we run a recursive feature selection (RFE) algorithm with a 10-fold cross-validation procedure. This has the goal of finding a subset of *features* that can be used to produce an accurate quality model. At each iteration of feature selection, the *features* are ranked. The less important *features* are sequentially eliminated. Figure 7.7 shows the selected *features* ranked by importance (shown in the Table 7.1).

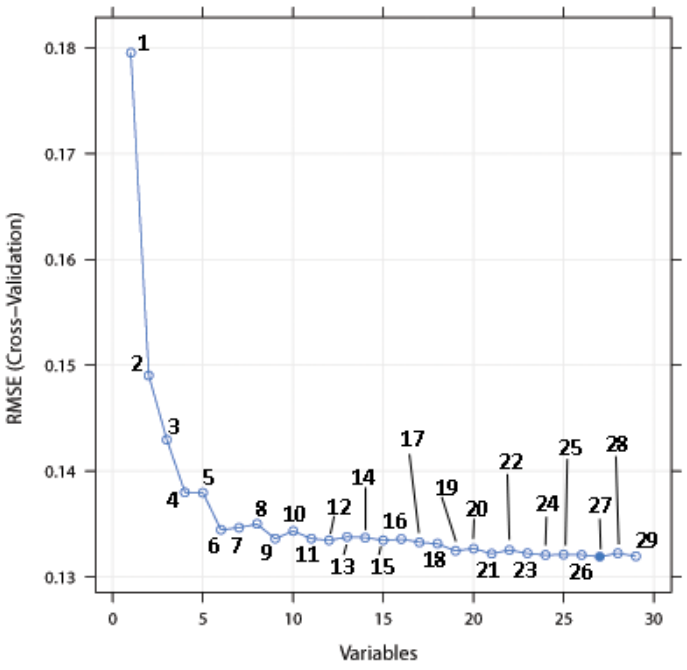


Figure 7.7: *Features* ranked by importance.

To compute the RFE algorithm for feature selection, we used the R v3.2.5 software with *rfe* function of the *caret* package. Since feature selection is part of the model building process, resampling methods (e.g. cross-validation) should factor in the variability caused by feature selection when calculating performance. So, we performed RFE algorithm incorporating resampling (k-fold cross-validation). In each resampling iteration, the data are partitioned into training and test/hold-back set via resampling, where the model is tuned/trained on the training set using all variables and the held-back samples are

Table 7.1: Selected *features* ranked by importance.

Rank	Features	Rank	Features	Rank	Features
1	$P_{intra_{16}}$	11	P_{intra_8}	21	$sumHorAC_8$
2	$P_{intra_{32}}$	12	$avgEnergyDC_{16}$	22	$sumAC_{16}$
3	$P_{inter_{32}}$	13	$sumEnergyDC_{32}$	23	$sumAC_8$
4	PV_V	14	$avgEnergyDC_8$	24	$sumVerAC_{16}$
5	P_{inter_8}	15	PV_H	25	$diffBorder_{32}$
6	$SalMap$	16	$avgEnergyDC_{32}$	26	$sumVerAC_{32}$
7	PS_V	17	$sumEnergyDC_{16}$	27	$sumHorAC_{32}$
8	PS_H	18	$diffBorder_8$	28	$diffBorders_{16}$
9	$P_{inter_{16}}$	19	$sumEnergyDC_8$	29	$sumHorAC_{16}$
10	$sumVerAC_8$	20	$sumAC_{32}$		

predicted. Next, the variable importance or ranking are computed. So, for each subset size ($S_i, i = 1 \dots S$), the S_i most important variable are kept and tune/train the model on the training set using S_i variables, by predicting the held-back samples. Next, the performance over the S_i are computed using the held-back samples and the appropriate number of variables are determined. The final list of variables to keep in the final model are estimated and, finally, the model is fitted based on the optimal S_i using the original training set.

Although the RFE algorithm shows that the best subset size has 27 *features*, for simplification, we used only the top 12 *features*: $P_{intra_{16}}$, $P_{intra_{32}}$, $P_{inter_{32}}$, PV_V, P_{inter_8}, $SalMap$, PS_V, PS_H, $P_{inter_{16}}$, $sumVerAC_8$, P_{intra_8}, and $avgEnergyDC_{16}$.[1] These selected *features* are the input to our non-linear SVR model, which combines them to obtain an overall quality estimate. Selecting only the top 12 *features* significantly decreased the training processing time, while maintaining a similar performance in terms of correlation coefficients.

7.3 Experimental Results

We tested the proposed approach using the test sequences of CSIQ, LIVE, IVPL, Exp.1a, Exp.2a, and Exp.3a databases. As mentioned earlier, the test consists of comparing the predicted and subjective scores, obtaining correlation coefficients. We compared our results with other traditional IQMs (see Table 2.2). Table 7.2 depicts these results. Notice

[1]The subscript numbers in the *features* indicate the size of the block from which it was extracted.

that the proposed metric performs better than most of the artifact metrics. The only exception is the results of $Blur_M$, $Blur_C$, and $Bloc_W$ for Exp.1a.

To examine how each metric responded to different types of distortions, we tested them on sequences of the CSIQ video database. Table 7.3 depicts the results separated by distortion types. Notice that, in general, all artifact metrics had a poor performance (correlation values lower than 0.140), whilst our method had a much higher performance, with correlation coefficients greater than 0.573. Tables 7.4 and 7.5 show the same results for the LIVE and IVPL video databases, respectively. Although the correlation values corresponding to the proposed metric are not high, they are still much higher than the values found for the artifact metrics (lower than 0.270). Finally, Tables 7.6, 7.7, and 7.8 show the same results for the Exp.1a, Exp.2a, and Exp.3a, respectively. For Exp.1a (only packet-loss), all metrics had similar performances, with the exception of $Blur_C$ that had a much higher performance. For Exp.2a and Exp.3a, the proposed metric achieved a much higher performance, with correlation coefficients greater than 0.830.

Table 7.2: Comparison of the correlation coefficients computed from set of video databases and artifact metrics.

Database	CSIQ		LIVE		IVPL		Exp.1a		Exp.2a		Exp.3a	
	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC
Proposed	**0.607**	**0.552**	**0.385**	**0.300**	**0.578**	**0.537**	0.462	0.471	**0.816**	**0.806**	**0.852**	**0.821**
$Pack_B$	0.081	0.085	-0.035	0.015	-0.012	0.042	0.352	0.318	-0.142	-0.108	0.072	-0.068
$Pack_R$	0.126	0.088	0.127	0.137	0.222	0.133	0.389	0.369	0.313	0.327	0.475	0.493
$Blur_M$	0.276	0.201	-0.024	-0.051	-0.120	-0.076	0.468	0.449	0.067	0.135	-0.016	-0.022
$Blur_N$	0.281	0.260	0.236	0.267	0.094	0.132	0.418	0.418	-0.047	-0.067	-0.121	-0.140
$Blur_C$	0.291	0.302	0.063	-0.014	0.129	0.066	**0.596**	**0.523**	0.291	0.165	0.216	0.282
$Bloc_F$	0.108	0.077	0.080	0.027	0.277	0.215	0.386	0.344	0.669	0.656	0.642	0.659
$Bloc_W$	0.190	0.121	0.124	0.049	0.175	0.200	0.486	0.480	-0.008	-0.022	0.107	0.124

7.3.1 Re-scaling the Data from Experiments

As mentioned in Chapter 4- Section 4.5, models scaled with INLSA, i.e. fitted on Re-scaled MAVs (RMAVs), obtained a better performance, showing that re-aligning the data before fitting the models is beneficial. Therefore, in this section, we used INLSA to test the same model with different video databases, which have different experimental methodologies. For example, in the IVPL database, the scores spread between 0 and 1, whilst in our experiments the scores corresponded to mean annoyance values ranging from

Table 7.3: Comparison of correlation coefficients per distortion in the CSIQ database.

Distortion	H.264		MJPEG		HEVC		Pack	
	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC
Proposed	**0.573**	**0.559**	**0.643**	**0.552**	**0.599**	**0.537**	**0.597**	**0.560**
$Pack_B$	0.067	0.060	0.049	0.105	0.043	0.036	0.028	0.071
$Pack_R$	-0.098	-0.124	-0.119	-0.178	-0.125	-0.164	-0.001	-0.057
$Blur_M$	0.089	0.071	0.148	0.172	0.046	0.094	0.048	0.013
$Blur_N$	0.177	0.113	0.150	0.148	0.176	0.153	0.127	0.074
$Blur_C$	0.131	0.125	-0.023	-0.121	0.157	0.110	0.128	0.076
$Bloc_F$	0.136	0.107	0.179	0.116	0.125	0.074	0.135	0.161
$Bloc_W$	-0.030	-0.079	0.063	-0.042	0.051	-0.023	0.129	0.060

Table 7.4: Comparison of correlation coefficients per distortion in the LIVE database.

Distortion	H.264		MPEG2		IP		Wireless	
	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC
Proposed	**0.477**	**0.430**	**0.502**	**0.439**	**0.480**	**0.414**	**0.523**	**0.451**
$Pack_B$	0.087	0.064	0.024	0.061	0.029	-0.002	0.027	0.015
$Pack_R$	0.172	0.197	0.162	0.180	0.106	0.157	0.077	0.107
$Blur_M$	-0.012	-0.019	0.084	0.030	0.030	0.032	-0.066	-0.076
$Blur_N$	0.261	0.245	0.263	0.241	0.214	0.179	0.267	0.238
$Blur_C$	0.171	0.168	0.057	0.044	0.064	0.060	0.050	0.037
$Bloc_F$	-0.040	-0.088	0.055	-0.004	-0.002	-0.023	0.017	-0.111
$Bloc_W$	0.025	-0.072	0.084	0.028	0.030	-0.017	0.081	0.062

0 and 100. Therefore, it is reasonable to assume that they may have spanned different ranges of quality that are not necessarily equivalent.

Before comparing the data from all experiments, we used INLSA to re-align the annoyance scores, using SSIM as the objective metric [29]. Also, since Experiment 3 has the highest number of artifact combinations, it was used as the reference experiment. Figures 7.8 show the MAV for the complete set of experiments before (top) and after (bottom) using INLSA, respectively, versus the corresponding SSIM value of the video. Notice that, after mapping the MAVs from all experiments onto the scale of Exp.3a, the MAVs of IVPL database spans a more comparable range of annoyance.

Table 7.5: Comparison of correlation coefficients per distortion in the IVPL database.

Distortion	H.264		MPEG2		IP	
	PCC	SCC	PCC	SCC	PCC	SCC
Proposed	**0.408**	**0.423**	**0.508**	**0.485**	**0.569**	**0.526**
$Pack_B$	0.097	0.132	0.137	0.159	0.064	0.151
$Pack_R$	0.111	0.020	0.298	0.164	0.114	0.114
$Blur_M$	-0.216	-0.219	-0.221	-0.206	-0.253	-0.247
$Blur_N$	0.078	0.103	0.146	0.139	0.151	0.173
$Blur_C$	0.131	0.124	0.108	0.080	0.086	0.023
$Bloc_F$	0.262	0.199	0.227	0.190	0.170	0.138
$Bloc_W$	0.110	0.136	0.178	0.172	0.105	0.181

Table 7.6: Comparison of correlation coefficients per distortion in the Exp.1a database.

	Distortion: packet-loss	
	PCC	SCC
Proposed	0.361	0.389
$Pack_B$	0.374	0.316
$Pack_R$	0.403	0.361
$Blur_M$	0.350	0.412
$Blur_N$	0.363	0.370
$Blur_C$	**0.619**	**0.529**
$Bloc_F$	0.478	0.429
$Bloc_W$	0.522	0.514

Taking a closer look at Figure 7.8, we can also notice that for the same SSIM values, each experiment has different MAVs (Figure 7.8 (a)). In particular, and as expected, for IVPL, the entire MAV range is clustered on the bottom part of the SSIM scale. Also, for Exp.1a, the MAV range is clustered on the top part of the SSIM scale. This means that videos with relatively low levels of impairments (as measured by SSIM) are judged as highly annoying. After mapping the MAVs, we can notice that the MAVs of IVPL span a more comparable range of annoyance. So, the re-scaled MAV (RMAV) allowing to analyze the data from all experiments as a whole.

We also used SVR to predict annoyance from all video databases using RMAVS. Table

Table 7.7: Comparison of correlation coefficients per distortion for Exp.2a database.

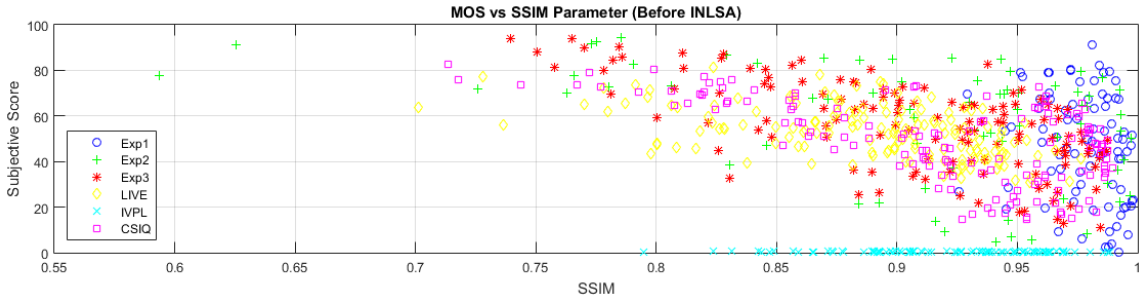
Distortion	bloc		blur		[bloc;blur]	
	PCC	SCC	PCC	SCC	PCC	SCC
Proposed	**0.830**	**0.804**	**0.859**	**0.848**	**0.822**	**0.829**
$Pack_B$	-0.247	-0.158	-0.057	0.001	-0.017	0.024
$Pack_R$	0.316	0.282	0.294	0.304	0.155	0.145
$Blur_M$	0.025	0.033	-0.041	-0.052	0.107	0.085
$Blur_N$	0.135	0.080	0.112	0.075	0.123	0.107
$Blur_C$	0.248	0.310	0.309	0.315	0.221	0.292
$Bloc_F$	0.696	0.721	0.715	0.748	0.660	0.703
$Bloc_W$	-0.223	-0.246	-0.190	-0.204	-0.230	-0.353

Table 7.8: Comparison of correlation coefficients per distortion in Exp.3a database.

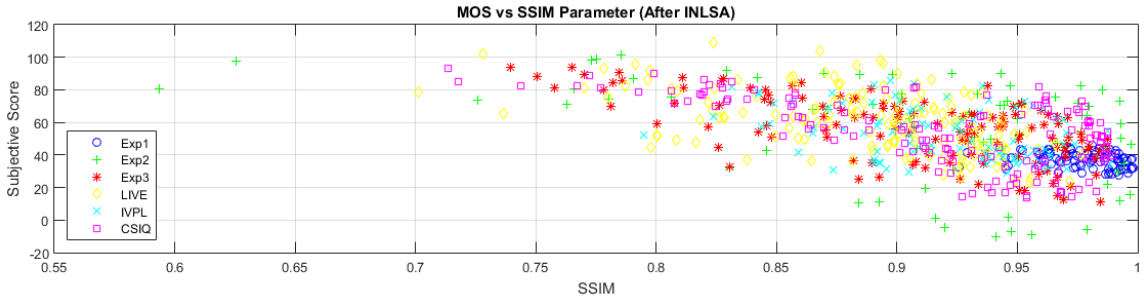
Distortion	bloc		blur		pdp		[pdp;bloc]		[pdp;blur]		[pdp;bloc;blur]	
	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC
Proposed	**0.860**	**0.831**	**0.864**	**0.842**	**0.875**	**0.835**	**0.865**	**0.864**	**0.849**	**0.819**	**0.858**	**0.864**
$Pack_B$	0.022	-0.050	0.158	0.075	0.037	-0.021	0.004	-0.001	0.009	-0.022	-0.037	0.022
$Pack_R$	0.401	0.485	0.389	0.421	0.328	0.401	0.322	0.384	0.377	0.453	0.360	0.410
$Blur_M$	-0.083	-0.062	0.002	-0.007	0.004	-0.053	-0.101	-0.118	-0.022	-0.039	-0.122	-0.183
$Blur_N$	-0.113	-0.123	-0.031	-0.089	-0.163	-0.096	-0.162	-0.182	0.012	0.054	-0.068	-0.019
$Blur_C$	0.267	0.239	0.232	0.223	0.233	0.207	0.251	0.242	0.232	0.156	0.209	0.220
$Bloc_F$	0.593	0.597	0.622	0.603	0.563	0.596	0.610	0.625	0.590	0.616	0.659	0.640
$Bloc_W$	-0.234	-0.289	-0.001	0.100	-0.187	-0.275	-0.199	-0.311	-0.194	-0.270	-0.086	-0.104

7.9 shows that the proposed metric using RMAV (PM_{RMAV}) has a better performance than using MAV (PM_{MAV}), with the exception of the data from Exp.1a. This results was expected since our *features* did not perform as well for videos impaired with only packet-loss according.

We also tested our method with three full-reference metrics: Gradient Magnitude Similarity Deviation (GMSD) [106], Structural Similarity Index (SSIM) [29], and Spatial and Spatiotemporal Slices via Gradient Magnitude Similarity Deviation (SSTS-GMSD) [107]. The results show that our metric performed a similar performance than both GMSD and SSIM metrics. When compared with SSTS-GMSD metric, our metric was better for all video databases, with the exception of CSIQ and Exp.1a.



(a)



(b)

Figure 7.8: (a) MAVs and (b) RMAVs (after applying INLSA [86]) versus SSIM for all Experiments.

Table 7.9: Comparison of the correlation coefficients computed from proposed metric (PM) using MAV and RMAVs.

Database	CSIQ		LIVE		IVPL		Exp.1a		Exp.2a		Exp.3a	
	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC
PM_{RMAV}	0.581	0.539	0.360	**0.355**	**0.895**	**0.884**	-0.037	0.035	**0.882**	**0.863**	**0.856**	0.804
PM_{MAV}	**0.607**	**0.552**	**0.385**	0.300	0.578	0.537	**0.462**	**0.471**	0.816	0.806	0.852	**0.821**

Table 7.10: Comparison of the correlation coefficients computed from PM_{RMAV} and VQMs using re-scaled MAVs.

Database	CSIQ		LIVE		IVPL		Exp.1a		Exp.2a		Exp.3a	
	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC
PM_{RMAV}	0.581	0.539	0.360	0.355	**0.895**	**0.884**	-0.037	0.035	**0.882**	**0.863**	**0.856**	**0.804**
GMSD	**0.808**	0.825	**0.729**	**0.726**	0.579	0.652	0.486	0.502	0.659	0.709	0.675	0.663
SSIM	-0.579	-0.541	-0.500	-0.525	-0.196	-0.204	**-0.534**	**-0.620**	-0.520	-0.584	-0.616	-0.599
SSTS-GMSD	0.795	**0.839**	-0.274	-0.290	0.396	0.487	0.413	0.405	0.663	0.729	0.719	0.702

7.4 Discussion

We compared the proposed method with 7 artifact metrics and 3 FR metrics, for 6 video quality databases. The proposed model (PM_{MAV}) uses a combination of artifact-based *features*, which were designed inspired by blockiness, blurriness, and packet-loss metrics. A non-linear SVR regression model is used to combine all *features* and generate an overall quality estimate. A comparison of the correlation coefficients for the 6 video quality databases showed that the proposed model performed better than the most of the artifact metrics. The only exception was the results obtained with $Blur_M$, $Blur_C$, and $Bloc_W$ for the Exp.1a. The proposed metric was also more effective in estimating the quality of isolated distortions in the public databases. When we re-scaled MAV using INLSA, the proposed metric (PM_{RMAV}) presented better results for IVPL, Exp.2a, and Exp.3a, although for Live and CSIQ the correlation coefficients were slightly different. Comparing the proposed algorithm with a set of FR metrics, the method (PM_{RMAV}) showed a good performance, specially considering that it does not require the reference.

Chapter 8

Conclusions and Future Works

8.1 Conclusions

We presented the results of six subjective experiments aimed at studying the characteristics of three artifacts (blockiness, blurriness, and packet-loss) commonly found in digital videos. The experiments had the goal of studying these artifacts and determining their relationship to quality, while investigating their interactions with each other. Results showed that annoyance increased with both the number of artifacts and their strength, with blockiness being the most annoying artifact. We proposed several models for predicting annoyance, including linear models with and without interactions and interception terms, Minkowski models, and a non-linear model based on SVR. Interactions were observed, suggesting that the overlap of multiple artifacts generated masking effects, what may decrease the annoyance perception. The correlation coefficients of fits re-scaled using INLSA (RMAVS) were higher than the fits using MAVs.

With respect to the appearance, visibility, and annoyance of these three artifacts, results showed that, when the artifact signals were presented alone at a high strength, participants were able to identify them correctly. At low strengths, on the other hand, other artifacts were reported. Annoyance increased with both the number of artifacts and their strength. We also proposed several models for predicting annoyance, where annoyance models were obtained by combining the artifact perceptual strengths (MSV), also using linear models with and without interactions and Minkowski models. Performing an ANOVA test, we found that all types of artifact signal strengths had a significant effect on MAV. Results also indicated that there were interactions between some of the artifact perceptual strengths. The non-linear model based on SVR provided greater values of correlation than the other tested models. In summary, results show that annoyance can be modeled as a multidimensional function of the individual artifact signal measurements.

We also studied the effect combinations of artifacts have on viewing behavior, by analyzing the viewing behavior of participants in terms of fixation duration and spatial gaze location. Results indicated that the presence of impairments has no impact on the duration of fixations, although differences in viewing behavior exist due to a change in task. Nevertheless, analyzing saliency maps we were able to detect changes in gaze deployment. In particular, we measured the similarity of saliency maps corresponding to the same video captured for different tasks (free-viewing or quality assessment), types of impairments (different combinations of packet-loss, blockiness and blurriness), or a combination of the two. Results showed that the presence of impairments in the video impacts the saliency distribution. We did not find an effect of a specific type of artifact combination on saliency changes.

Interestingly, the similarity measure of the saliency maps increased with the increase of the artifact annoyance. This is a counter intuitive result, as one would expect more annoying artifacts to be visually stronger and thus create saliency on their own. A possible explanation for this result is that, for combinations with low MAV the impact of localized packet-loss artifacts is more evident. But, for more annoying combinations (higher MAVs) this localized effect may be masked by the presence of other artifacts. So, the source of annoyance may become indistinctively diffuse across the whole video area. For this reason, further analysis is needed to link the change in saliency to physical properties of the video.

Finally, we proposed a no-reference metric that estimates the quality of video using artifact-based *features*. The *features* are extracted aiming to characterize the 3 artifacts: blockiness, blurriness, and packet-loss, and they are combined using a non-linear SVR regression algorithm. The proposed metric was tested using 7 traditional artifact metrics and 3 full-reference metrics. In general, our proposed metric showed a good overall performance over other metrics. In particular, over traditional artifact metrics. Also, our metric had better performance than full-reference metrics in IVPL, Exp.2a, and Exp.3a (see Table 7.10). These results are very interesting since it is expected that FR metrics take advantage over NR metrics.

8.2 Future Works

In Chapter 5, we found that the 3 types of artifact signal strengths had a significant effect on MAV, with blockiness having the strongest effect. Also, our annoyance models indicated that there are interactions among some of the artifact perceptual strengths. Also, it would be important to validate our annoyance models by testing these models in naturally occurring artifacts. Unfortunately, we were not able to perform these tests because the available artifact metrics are not robust enough to provide a reliable measure

of the artifact combinations. Also, the currently available video quality databases do not have a diverse number of distortions in combinations, which is needed by this type of work.

Another possible future work is to improve the design of the proposed no-reference video quality metric. In this work, we only took into consideration features related to the artifacts blockiness, blurriness, and packet-loss. However, in real-world scenarios, additional types of artifacts can also be present. Examples of other types of artifacts currently present in digital videos include quantization noise, ringing, color distortions, contrast distortions, jitter, etc. Therefore, it would be interesting to determine what features can be used to detect and estimate the strength of these additional artifacts,.

Given that an open problem in image quality is the quality assessment of enhanced or high quality images, the proposed approach could also incorporate *features* inspired by intrinsic characteristics of the image that directly affect quality, like color, naturalness, sharpness, contrast, etc. This issue is particularly important in the current digital video scenario, which makes possible to transmit video at a very high quality (4K, HDR, etc.).

Given the current state-of-the-art of the visual-attention computational models, we believe the proposed no-reference video quality metric can also be improved by including additional visual attention features. For example, the metric can include temporal saliency *features* and top-down visual attention aspects. Also, the knowledge about the saliency of different types of distortions can be used to study and interaction model.

8.3 Publications

(SPIE, 2015) A. F. Silva, M. C. Farias, J. A. Redi, Assessing the influence of combinations of blockiness, blurriness, and packet-loss impairments on visual attention deployment, in: IS&T/SPIE Electronic Imaging, Vol. 9394, 2015, pp. 93940Z-93940Z-11.

(QoMex, 2015) P. G. Freitas, J. A. Redi, M. C. Farias, A. F. Silva, Video quality ruler: A new experimental methodology for assessing video quality. In: 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), 2015, Pilos. pp. 1-6.

(QoMex, 2016) A. F. Silva, M. C. Farias, J. A. Redi, Annoyance models for videos with spatio-temporal artifacts, In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, 2016.

(SBRT, 2016) D. D. R. Morais, A. F. Silva, M. C. Q. Farias, A Correlation-Based No-Reference Packet-Loss Metric, In: XXXIV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais - SBrT2016, August 30 to September 02, Santarém, PA.

(TMM, 2106) A. F. Silva, M. C. Farias, J. A. Redi, Perceptual annoyance models for videos with combinations of spatial and temporal artifacts, IEEE Transactions on Multimedia, DOI:10.1109/TMM.2016.2601027.

Bibliography

[1] Brian Keelan. *Handbook of image quality: characterization and prediction*. CRC Press, 2002. 1

[2] Michael Yuen and HR Wu. A survey of hybrid mc/dpcm/dct video coding distortions. *Signal processing*, 70(3):247–278, 1998. 1

[3] ITU RDS. Methodology for the subjective assessment of the quality of television pictures. *ITU-R BT*, pages 500–1, 2002. 1, 6, 15

[4] Alexandre F Silva, Mylène CQ Farias, and Judith A Redi. Assessing the influence of combinations of blockiness, blurriness, and packet loss impairments on visual attention deployment. In *SPIE/IS&T Electronic Imaging*, pages 93940Z–93940Z. International Society for Optics and Photonics, 2015. 1, 9, 42

[5] Judith A Redi. Visual quality beyond artifact visibility. In *IS&T/SPIE Electronic Imaging*, pages 86510N–86510N. International Society for Optics and Photonics, 2013. 1

[6] Stefan Winkler and Christof Faller. Perceived audiovisual quality of low-bitrate multimedia content. *IEEE transactions on multimedia*, 8(5):973–980, 2006. 1, 4

[7] Weisi Lin and C-C Jay Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312, 2011. 1, 3, 4, 8

[8] Anush Krishna Moorthy and Alan Conrad Bovik. Visual quality assessment algorithms: what does the future hold? *Multimedia Tools and Applications*, 51(2):675–696, 2011. 1, 2, 3, 4

[9] Junyong You, Jari Korhonen, Andrew Perkis, and Touradj Ebrahimi. Balancing attended and global stimuli in perceived video quality assessment. *IEEE Transactions on Multimedia*, 13(6):1269–1285, 2011. 1

[10] Mylène CQ Farias and Sanjit K Mitra. Perceptual contributions of blocky, blurry, noisy, and ringing synthetic artifacts to overall annoyance. *Journal of Electronic Imaging*, 21(4):043013–043013, 2012. 1, 2, 35, 38, 41, 62, 63

[11] Mylene Christine Queiroz de Farias. *No-reference and reduced reference video quality metrics: new contributions*. University of California, Santa Barbara, 2004. 1, 7, 73

[12] Margaret H Pinson and Stephen Wolf. An objective method for combining multiple subjective data sets. In *VCIP*, pages 583–592, 2003. 2, 32, 33

[13] R Venkatesh Babu, Ajit S Bopardikar, Andrew Perkis, and Odd Inge Hillestad. No-reference metrics for video streaming applications. In *International Workshop on Packet Video*, 2004. 2, 10, 11

[14] Mylene CQ Farias and Sanjit K Mitra. No-reference video quality metric based on artifact measurements. In *IEEE International Conference on Image Processing 2005*, volume 3, pages III–141. IEEE, 2005. 2, 3, 4, 10, 11, 63, 78

[15] Hantao Liu and Ingrid Heynderickx. A perceptually relevant no-reference blockiness metric based on local image characteristics. *EURASIP Journal on Advances in Signal Processing*, 2009(1):1–14, 2009. 2, 9, 42

[16] Jorge Caviedes and Joel Jung. No-reference metric for a video quality control loop. In *Proceedings of 5th World Multiconference on Systemics, Cybernetics, and Informatics*, pages 290–295, 2001. 2, 3, 9, 10, 42

[17] Vishwakumara Kayargadde and Jean-Bernard Martens. Perceptual characterization of images degraded by blur and noise: model. *JOSA A*, 13(6):1178–1188, 1996. 2, 42

[18] Huib De Ridder. Minkowski-metrics as a combination rule for digital-image-coding impairments. In *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*, pages 16–26. International Society for Optics and Photonics, 1992. 2, 38, 41, 42, 62

[19] Damon M Chandler, Kenny H Lim, and Sheila S Hemami. Effects of spatial correlations and global precedence on the visual fidelity of distorted images. In *Electronic Imaging 2006*, pages 60570F–60570F. International Society for Optics and Photonics, 2006. 2

[20] Mylne CQ Farias, John M Foley, and Sanjit K Mitra. Detectability and annoyance of synthetic blocky, blurry, noisy, and ringing artifacts. *IEEE Transactions on Signal Processing*, 55(6):2954–2964, 2007. 2, 13

[21] Michael S Moore, John M Foley, and Sanjit K Mitra. Defect visibility and content importance: effects on perceived impairment. *Signal Processing: Image Communication*, 19(2):185–203, 2004. 2

[22] Quan Huynh-Thu and Mohammed Ghanbari. Modelling of spatio-temporal interaction for video quality assessment. *Signal Processing: Image Communication*, 25(7):535–546, 2010. 2

[23] Amy R Reibman, Vinay A Vaishampayan, and Yegnaswamy Sermadevi. Quality monitoring of video over a packet network. *IEEE Transactions on Multimedia*, 6(2):327–334, 2004. 2

[24] Guangtao Zhai, Jianfei Cai, Weisi Lin, Xiaokang Yang, Wenjun Zhang, and Minoru Etoh. Cross-dimensional perceptual quality assessment for low bit-rate videos. *IEEE Transactions on Multimedia*, 10(7):1316–1324, 2008. 2

[25] Matteo Naccari, Marco Tagliasacchi, and Stefano Tubaro. No-reference video quality monitoring for h. 264/avc coded video. *IEEE Transactions on Multimedia*, 11(5):932–946, 2009. 2

[26] Ulrich Engelke. Modelling perceptual quality and visual saliency for image and video communications. 2010. 3

[27] Mylène CQ Farias, I Heynderickx, BL Macchiavello Espinoza, and JA Redi. Visual artifacts interference understanding and modeling (varium). In *Seventh international workshop on video processing and quality metrics for consumer electronics*, volume 1, 2013. 3, 8, 10

[28] Scott J Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. In *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*, pages 2–15. International Society for Optics and Photonics, 1992. 4, 9

[29] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4, 9, 23, 33, 86, 88

[30] David S Hands. A basic multimedia quality model. *IEEE Transactions on multimedia*, 6(6):806–816, 2004. 4

[31] Zhou Wang, Hamid R Sheikh, and Alan C Bovik. No-reference perceptual quality assessment of jpeg compressed images. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–477. IEEE, 2002. 4, 9, 10, 42

[32] Itu-t recommendation p.930: Principles of a reference impairment system for video. 1996. 6, 14, 15

[33] Sander Sunde Vorren. Subjective quality evaluation of the effect of packet loss in high-definition video. 2006. 7

[34] Alexandre F Silva, Mylene Farias, and Judith A Redi. Perceptual annoyance models for videos with combinations of spatial and temporal artifacts. *IEEE Transactions on Multimedia*. 7, 13, 14, 15, 21, 62, 74

[35] Bernd Girod. What's wrong with mean-squared error? In *Digital images and human vision*, pages 207–220. MIT press, 1993. 8, 73

[36] Zhou Wang and Alan C Bovik. Mean squared error: love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009. 8, 73

[37] Miguel O Martínez-Rach, Pablo Piñol, Otoniel M López, Manuel Perez Malumbres, José Oliver, and Carlos Tavares Calafate. On the performance of video quality assessment metrics under different compression and packet loss scenarios. *The Scientific World Journal*, 2014, 2014. x, 8

[38] Jeffrey Lubin. The use of psychophysical data and models in the analysis of display system performance. In *Digital images and human vision*, pages 163–178. MIT Press, 1993. 9

[39] Jeffrey Lubin. A visual discrimination model for imaging system design and evaluation. *Vision models for target detection and recognition*, 2:245–357, 1995. 9

[40] Margaret H Pinson and Stephen Wolf. A new standardized method for objectively measuring video quality. *IEEE Transactions on broadcasting*, 50(3):312–322, 2004. 9

[41] Irwan Prasetya Gunawan and Mohammed Ghanbari. Image quality assessment based on harmonics gain/loss information. In *IEEE International Conference on Image Processing 2005*, volume 1, pages I–429. IEEE, 2005. 9

[42] Mathieu Carnec, Patrick Le Callet, and Dominique Barba. An image quality assessment method based on perception of structural information. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 3, pages III–185. IEEE, 2003. 9

[43] SD Voran and Stephen Wolf. The development and evaluation of an objective video quality assessment system that emulates human viewing panels. In *Broadcasting Convention, 1992. IBC., International*, pages 504–508. IET, 1992. 9

[44] Stephen Wolf and Margaret H Pinson. Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system. In *Proc. SPIE*, volume 3845, pages 266–277, 1999. 9

[45] HR Wu and M Yuen. A generalized block-edge impairment metric for video coding. *IEEE Signal Processing Letters*, 4(11):317–320, 1997. 9, 10, 63

[46] Zhou Wang, Alan C Bovik, and BL Evan. Blind measurement of blocking artifacts in images. In *Image Processing, 2000. Proceedings. 2000 International Conference on*, volume 3, pages 981–984. Ieee, 2000. 9, 10, 11, 63

[47] Cristina Oprea, Ionut Pirnog, Constantin Paleologu, and Mihnea Udrea. Perceptual video quality assessment based on salient region detection. In *Telecommunications, 2009. AICT'09. Fifth Advanced International Conference on*, pages 232–236. IEEE, 2009. 9

[48] Donald Bailey, Marco Carli, Mylene Farias, and Sanjit Mitra. Quality assessment for block-based compressed images and videos with regard to blockiness artifacts. In *International Workshop in Data Compression*, 2002. 9

[49] Ville Ojansivu, Olli Silvén, and Risto Huotari. A technique for digital video quality evaluation. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 3, pages III–181. IEEE, 2003. 9

[50] Alexander KG Wörner. A real time single ended algorithm for objective quality monitoring of compressed video signals. In *Advanced Motion Imaging Conference, 36th SMPTE Annual*, pages 1–8. SMPTE, 2002. 9

[51] R Venkatesh Babu, Andrew Perkis, and Odd Inge Hillestad. Evaluation and monitoring of video quality for uma enabled video streaming systems. *Multimedia Tools and Applications*, 37(2):211–231, 2008. 9, 42

[52] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi. Perceptual blur and ringing metrics: application to jpeg2000. *Signal processing: Image communication*, 19(2):163–172, 2004. 9, 10, 11, 42

[53] T Vlachos. Detection of blocking artifacts in compressed video. *Electronics Letters*, 36(13):1106–1108, 2000. 10

[54] Niranjan D Narvekar and Lina J Karam. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *IEEE Transactions on Image Processing*, 20(9):2678–2683, 2011. 10, 11

[55] Frederique Crete, Thierry Dolmiere, Patricia Ladret, and Marina Nicolas. The blur effect: perception and estimation with a new no-reference perceptual blur metric. In *Electronic Imaging 2007*, pages 64920I–64920I. International Society for Optics and Photonics, 2007. 10, 11, 79, 80

[56] Hua-xia Rui, Chong-rong Li, and Sheng-ke Qiu. Evaluation of packet loss impairment on streaming video. *Journal of Zhejiang University SCIENCE A*, 7(1):131–136, 2006. 10, 11

[57] Cuong T Vu, Eric C Larson, and Damon M Chandler. Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience. In *Image Analysis and Interpretation, 2008. SSIAI 2008. IEEE Southwest Symposium on*, pages 73–76. IEEE, 2008. 11, 64

[58] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995. 11, 64

[59] Junle Wang, Damon M Chandler, and Patrick Le Callet. Quantifying the relationship between visual salience and visual importance. In *IS&T/SPIE Electronic Imaging*, pages 75270K–75270K. International Society for Optics and Photonics, 2010. 11, 64

[60] Judith Redi, Hantao Liu, Rodolfo Zunino, and Ingrid Heynderickx. Interactions of visual attention and quality perception. In *IS&T/SPIE Electronic Imaging*, pages 78650S–78650S. International Society for Optics and Photonics, 2011. 11, 12, 64

[61] Ulrich Engelke, Hagen Kaprykowsky, Hans-Jürgen Zepernick, and Patrick Ndjiki-Nya. Visual attention in quality assessment. *IEEE Signal Processing Magazine*, 28(6):50–59, 2011. 11, 22, 64

[62] Judith Redi, Hantao Liu, Paolo Gastaldo, Rodolfo Zunino, and Ingrid Heynderickx. How to apply spatial saliency into objective metrics for jpeg compressed images? In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 961–964. IEEE, 2009. 11, 64

[63] Olivier Le Meur and Patrick Le Callet. What we see is most likely to be what matters: Visual attention and applications. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 3085–3088. IEEE, 2009. 11, 65

[64] Hani Alers, Judith Redi, Hantao Liu, and Ingrid Heynderickx. Studying the effect of optimizing image quality in salient regions at the expense of background content. *Journal of Electronic Imaging*, 22(4):043012–043012, 2013. 11, 65

[65] Alexandre Ninassi, Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Arnaud Tirel. Task impact on the visual attention in subjective image quality assessment. In *Signal Processing Conference, 2006 14th European*, pages 1–5. IEEE, 2006. 11, 22, 65

[66] Judith A Redi and Ingrid Heynderickx. Image quality and visual attention interactions: towards a more reliable analysis in the saliency space. In *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*, pages 201–206. IEEE, 2011. 12, 23

[67] Olivier Le Meur, Alexandre Ninassi, Patrick Le Callet, and Dominique Barba. Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric. *Signal Processing: Image Communication*, 25(7):547–558, 2010. 12, 65

[68] Judith Redi, Ingrid Heynderickx, Bruno Macchiavello, and Mylene Farias. On the impact of packet-loss impairments on visual attention mechanisms. In *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*, pages 1107–1110. IEEE, 2013. 12, 13, 22, 24, 25, 65, 66, 67, 69

[69] Claire Mantel, Nathalie Guyader, Patricia Ladret, Gelu Ionescu, and Thomas Kunlin. Characterizing eye movements during temporal and global quality assessment of h. 264 compressed video sequences. In *IS&T/SPIE Electronic Imaging*, pages 82910Y–82910Y. International Society for Optics and Photonics, 2012. 12, 65

[70] Olivier Le Meur, Alexandre Ninassi, Patrick Le Callet, and Dominique Barba. Do video coding impairments disturb the visual attention deployment? *Signal Processing: Image Communication*, 25(8):597–609, 2010. 12, 65, 68

[71] Ulrich Engelke, Romuald Pepion, Patrick Le Callet, and Hans-Jürgen Zepernick. Linking distortion perception and visual saliency in h. 264/avc coded video containing packet loss. In *Visual Communications and Image Processing 2010*, pages 774406–774406. International Society for Optics and Photonics, 2010. 12, 65

[72] Video Quality Experts Group (VQEG). Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase i. Technical report, Video Quality Experts Group (VQEG), 2008. 13

[73] A Ostaszewska and R Kłoda. Quantifying the amount of spatial and temporal information in video test sequences. In *Recent Advances in Mechatronics*, pages 11–15. Springer, 2007. 13

[74] Varium project video database. `http://www.ene.unb.br/mylene/databases.htm`. Accessed: 2016-01-30. 13

[75] Mylène CQ Farias, John M Foley, and Sanjit K Mitra. Perceptual analysis of video impairments that combine blocky, blurry, noisy, and ringing synthetic artifacts. In *Electronic Imaging 2005*, pages 107–118. International Society for Optics and Photonics, 2005. 15

[76] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. Study of subjective and objective quality assessment of video. *IEEE transactions on image processing*, 19(6):1427–1441, 2010. 19

[77] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan C Bovik, and Lawrence K Cormack. A subjective study to evaluate video quality assessment algorithms. In *IS&T/SPIE Electronic Imaging*, pages 75270H–75270H. International Society for Optics and Photonics, 2010. 19

[78] Phong V Vu and Damon M Chandler. Vis3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electronic Imaging*, 23(1):013016–013016, 2014. 19

[79] Alexandre F. Silva, Mylène C.Q. Farias, and Judith A. Redi. Annoyance models for videos with spatio-temporal artifacts. In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, 2016. 21, 62, 63

[80] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998. 22, 39

[81] Payam Refaeilzadeh, Lei Tang, and Huan Liu. Cross-validation. In *Encyclopedia of database systems*, pages 532–538. Springer, 2009. 22

[82] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987. 22

[83] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Saliency and human fixations: state-of-the-art and study of comparison metrics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1153–1160, 2013. 23

[84] Fadi Boulos, Benoît Parrein, Patrick Le Callet, and David Hands. Perceptual effects of packet loss on h. 264/avc encoded videos. In *Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM-09*, 2009. 25

[85] Huib de Ridder. Cognitive issues in image quality measurement. *Journal of Electronic Imaging*, 10(1):47–55, 2001. 32

[86] Stephen D Voran. *An iterated nested least-squares algorithm for fitting multiple data sets*. US Department of Commerce, National Telecommunications and Information Administration, 2002. x, xi, 33, 34, 89

[87] Antonis Papadogiannakis, Alexandros Kapravelos, Michalis Polychronakis, Evangelos P Markatos, and Augusto Ciuffoletti. Passive end-to-end packet loss estimation for grid traffic monitoring. In *Proceedings of the CoreGRID Integration Workshop*, pages 79–93, 2006. 34

[88] Ahmad Vakili and Jean-Charles Grégoire. Estimation of packet loss probability from traffic parameters for multimedia over ip. In *Proc. of the Seventh International Conference on Networking and Services, ICNS*, pages 44–48, 2011. 34

[89] Paolo Gastaldo, Rodolfo Zunino, and Judith Redi. Supporting visual quality assessment with machine learning. *EURASIP Journal on Image and Video Processing*, 2013(1):1–15, 2013. 37, 39, 63

[90] Albert J Ahumada and Cynthia H Null. Image quality: A multidimensional problem. *Digital images and human vision*, pages 141–148, 1993. 42

[91] MRM Nijenhuis and FJJ Blommaert. Perceptual error measure for sampled and interpolated images. *Journal of Imaging Science and Technology*, 41(3):249–258, 1997. 42

[92] Huib de Ridder and Gijberta M Majoor. Numerical category scaling: an efficient method for assessing digital image coding impairments. In *SC-DL tentative*, pages 65–77. International Society for Optics and Photonics, 1990. 42

[93] Mylène CQ Farias and Welington YL Akamine. On performance of image quality metrics enhanced with visual attention computational models. *Electronics letters*, 48(11):631–633, 2012. 64

[94] Patrick Le Callet, Stéphane Péchard, Sylvain Tourancheau, Alexandre Ninassi, and Dominique Barba. Towards the next generation of video and image quality metrics: Impact of display, resolution, contents and visual attention in subjective assessment. In *Second International Workshop on Image Media Quality and its Applications, IMQA2007*, page A2, 2007. 66

[95] Mario Vranješ, Snježana Rimac-Drlje, and Krešimir Grgić. Review of objective video quality metrics and performance comparison using different databases. *Signal Processing: Image Communication*, 28(1):1–19, 2013. 73

[96] Hong Ren Wu and Kamisetty Ramamohan Rao. *Digital video image quality and perceptual coding*. CRC press, 2005. 73

[97] Shyamprasad Chikkerur, Vijay Sundaram, Martin Reisslein, and Lina J Karam. Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE Transactions on Broadcasting*, 57(2):165–182, 2011. 73

[98] Xavier Marichal, Wei-Ying Ma, and HongJiang Zhang. Blur determination in the compressed domain using dct information. In *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, volume 2, pages 386–390. IEEE, 1999. 74

[99] Jorge Caviedes and Sabri Gurbuz. No-reference sharpness metric based on local edge kurtosis. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 3, pages III–53. IEEE, 2002. 74

[100] Atsuro Ichigaya, Yukihiro Nishida, and Eisuke Nakasu. Nonreference method for estimating psnr of mpeg-2 coded video by using dct coefficients and picture energy. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(6):817–826, 2008. 74

[101] K Bhattacharyya and HS Jamadagni. Dct coefficient-based error detection technique for compressed video stream. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 3, pages 1483–1486. IEEE, 2000. 74, 75

[102] Wei Zhang, Yi Tian, Xiaojie Zha, and Hantao Liu. Benchmarking state-of-the-art visual saliency models for image quality assessment. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1090–1094. IEEE, 2016. 82

[103] Lin Zhang, Ying Shen, and Hongyu Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, 2014. 82

[104] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pages 545–552. MIT Press, 2006. 82

[105] Laurent Itti, Christof Koch, Ernst Niebur, et al. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998. 82

[106] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, 2014. 88

[107] Peng Yan, Xuanqin Mou, and Wufeng Xue. Video quality assessment via gradient magnitude similarity deviation of spatial and spatiotemporal slices. In *SPIE/IS&T Electronic Imaging*, pages 94110M–94110M. International Society for Optics and Photonics, 2015. 88