DISSERTAÇÃO DE MESTRADO

A Hybrid No-Reference Video Quality Metric for Digital Transmission Applications

Dário Daniel Ribeiro Morais

Brasília, Março de 2017

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA
Faculdade de Tecnologia

DISSERTAÇÃO DE MESTRADO

A Hybrid No-Reference Video Quality Metric for Digital Transmission Applications

Dário Daniel Ribeiro Morais

Dissertação de Mestrado submetida ao Departamento de Engenharia Elétrica como requisito parcial para obtenção do grau de Mestre em Engenharia de Sistemas Eletrônicos e Automação

Banca Examinadora

Prof. Dr. Mylène Christine Queiroz de Farias _____
Orientador

Prof. Dr. Francisco Assis de Oliveira Nascimento _____
Examinador interno

Prof. Dr. Bruno Luiggi Macchiavello Espinoza _____
Examinador externo

FICHA CATALOGRÁFICA

MORAIS, DARIO
A Hybrid No-Reference Video Quality Metric for Digital Transmission Applications [Distrito Federal] 2017.
xvi, 47 p., 210 x 297 mm (ENE/FT/UnB, Mestre, Engenharia Elétrica, 2017).
Dissertação de Mestrado - Universidade de Brasília, Faculdade de Tecnologia.
Departamento de Engenharia Elétrica

1. Packet loss	2. Hybrid Metric
3. Video quality	4. Blockiness, Bluriness
I. ENE/FT/UnB	II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

MORAIS, D. (2017). *A Hybrid No-Reference Video Quality Metric for Digital Transmission Applications*. Dissertação de Mestrado, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 47 p.

CESSÃO DE DIREITOS

AUTOR: Dário Daniel Ribeiro Morais
TÍTULO: A Hybrid No-Reference Video Quality Metric for Digital Transmission Applications.
GRAU: Mestre em Engenharia de Sistemas Eletrônicos e Automação ANO: 2017

Dário Daniel Ribeiro Morais
Depto. de Engenharia Elétrica (ENE) - FT
Universidade de Brasília (UnB)
Campus Darcy Ribeiro
CEP 70919-970 - Brasília - DF - Brasil

Acknowledgments

Life is a marvellous experience. Discipline, passion, enthusiasm, courage, self confidence, persistence, hardwork and love are some of the elements that cannot be missed in the way. These features are also essentials to develop a long term project with lots of ups and downs, disciplines, study, articles, etc. All in the process is important, from the moment you decide to apply your documents to final stage of writing the dissertation or even in the point of presenting your work. Everything matters.

I definitely cannot forget the great people I have met and the fantastic moments we have shared together (People from all over Brazil and some countries of the earth globe). I could not have decided a better city to do my Master Course. Brasilia is an amazing city! Its diversity, politics, weather, culture, social life are some interesting aspects of this city. Likewise, the locals received in a quite kind and cozy way and they have been very important in my life. They have given me a considerable support and It has allowed me to settle down faster in the Capital.

Although the great moments have overcome the hard times, nothing in life comes without a hard effort and in my case this has not been different. For more than 2 years I have been dedicating my mind, work and soul for achieving the desirable Master's degree. There has been difficult times. Moments when you feel anxious, feel afraid of facing big challenges. But, all of this is part of the process of learning, reaching new knowledge to go further. Austerity times make the human being stronger and if we learn from it we can go anywhere.

Firsly, I would like to thank God for this wonderful opportunity, my family and the great people I have met along my way. I also want to thank professor Mylène for her technical support and all patience and dedication to me and all her postgraduate students. Moreover, I need to say a BIG Thank you to my family who has supported me unconditionally. My mother (Mariza) for always believing in education, my doctor brother (Baltazar) for his wise tips and my sister and brother-in-law (Mariana and Renan) for the unconditional support in Brasilia.

Next, I could not forget my friends and colleagues who have been with me in this journey. More specifically, Mélanie Robin, Nuria Bernardez, Kleber Saúde, Juliano Silveria, Gustavo Sandri, Vinicius Oliveira, Welington Akamine, Danilo Amaral, Gizele Abdon, Fadhil Firyaguna, Lucas, Jonathan Alis, Pedro Garcia, Helard Becerra, Elton Sarmanho and Daniel Souza.

Finally, Thanks Daisy Oliveira for the coaching support. The decision of studying in Brasilia came out inside the coach process. Irineu and Chiquinho Alves for the dance monitorship at Bycia. Pedro Mariano and Ana Amélia Diniz for enjoyable moments in the south american culture events and la Salsa, Bachata, Merengue and Reggaeton. Alexandre Fieno for our partnership during the Master course.

To finalize I would like to leave an encouragement message that I learned when I lived in England and it follows me wherever I go in life: Where there is a will, there is a way.

RESUMO

Este trabalho visa desenvolver uma métrica híbrida de qualidade de vídeo sem referência para aplicações de transmissão digital, que leva em consideração três tipos de artefatos: perda de pacotes, blocado e borrado. As características desses artefatos são extraídas a partir das sequências de vídeo a fim de quantificar a força desses artefatos. A avaliação de perda de pacotes é dividida em 2 etapas: detecção e medição. As avaliações de blocado e borrado seguem referências da literatura. Depois de obter as características dos três tipos de artefatos, um processo de aprendizado de máquina (SVR) é utilizado para estimar a nota de qualidade prevista a partir das características extraídas.

Os resultados obtidos com a métrica proposta foram comparados com os resultados obtidos com outras três métricas disponíveis na literatura (duas métricas NR de perda de pacotes e 1 métrica FR) e eles são promissores. A métrica proposta é cega, rápida e confiável para ser usada em cenários em tempo real.

ABSTRACT

This work aims to develop a hybrid no-reference video quality metric for digital transmission applications, which takes into account three types of artifacts: packet-loss, blockiness and bluriness. Features are extracted from the video sequences in order to quantity the strength of these three artifacts. The assessment of the packet-loss strength is performed in 2 stages: detection and measurement. The assessment of the strength of blockiness and blussiness follow references from literature. After obtaining the features from these three types of artifacts, a machine learning algorithm (the support vector regression technique), is used to estimate the predicted quality score from the extracted features.

The results obtained with the proposed metric were compared with the results obtained with three other metrics available in the literature (two NR packet-loss metrics and one FR metric). The proposed metric is blind, fast, and reliable to be used in real-time scenarios.

CONTENTS

Figure List

Table List

Acronyms

CSIQ	Computational and subjective image quality
DCT	Discrete Cosine Transform
FPS	Frame por second
FR	Full reference
GOP	Group of pictures
HEVC	High efficiency video coding
HRC	Hypothetical Reference Circuits
HVS	Human visual system
IP	Internet protocol
IQA	Image quality assessment
ITU	International Telecommunication Union
IVPL	Image and video processing laboratory
JPEG	Joint photographic experts group
MB	Macroblock
MOS	Mean opnion score
MAV	Mean annoyance values
MJPEG	Motion joint photographic experts group
MPEG	Moving picture experts group
MSE	Mean square error
NF	Number of frames
NR	No reference
PCC	Pearson correlation coefficient
PSNR	Peak signal-to-noise ratio
QoS	Quality of service
QoE	Quality of experience
RR	Reduced reference
SAD	Sum of absolute differences
SCC	Spearman correlation coefficient
SSIM	Structural similarity
SVM	Support vector machine
SVR	Support vector regression
Varium	Visual artifacts varium interference understanding and modeling
VQEG	Video quality experts group

1 INTRODUCTION

The consumption of internet services has grown sharply worldwide, especially over the last 10 years. The internet has become an essential tool in modern life. Social networks have conquered considerable space and most people have Facebook, Instagram or Twitter accounts. The use of smartphones and tablets and their applications have also proliferated. The applications are the most diverse possible, like online chat, browsers, maps, music readers, radio and games. Due to the wide use of smartphones, social network and high speed connections, video consumption has increased and it now corresponds to 80 percent of the total bandwidth used worldwide [1]. As a consequence, the need to assess the quality of videos that travels over the network has also increased over the last years.

Subjective quality assessment methods are considered the most precise way of estimating video quality [2]. These methods consist of performing experiments in which non-experts human participants (usually 15 - 30) are asked to rate the quality of a set of test videos. To ensure reproducibility and precision, these experiments are performed in controlled environments, following recommendations by International Telecommunication Union (ITU) [3]. For each test video, the average of the scores given by all participants, i.e. Mean Opinion Score (MOS), provides an estimate of the quality of that video as perceived by human observers. MOS values are generally used as a benchmark to test objective video quality metrics, which are basically algorithms that estimate video quality by making physical measurements of the signal.

Depending on the amount of information that is required at the measuring point, objective metrics can be classified into three categories: Full Reference (FR), Reduced Reference (RR) and No-Reference (NR) methods. Figure 1.1 illustrates these three types of metrics. Notice that, FR methods perform a comparison between original and test videos. This means that both test (received) and reference (original) videos must be available at the measuring point. For RR methods, some characteristics of the original video are extracted at the sender and transmitted to the receiver. At the receiver, the characteristics of the test video are acquired and compared to the characteristics of the original video. Although RR methods do not require the full original, they still need partial information about it. NR methods estimate the quality of a video without requiring the original. Since the original videos are not readily available in real-time scenarios, such as wireless communications, video over IP application, cable TV, Digital TV, etc. The development of NR methods is a rather important area of research.

As mentioned previously, the most accurate way to determine video quality is by measuring it using psychophysical experiments with human subjects. Unfortunately, these experiments are expensive, time-consuming and hard to incorporate into an automatic system. Therefore, there is currently a need for fast and accurate algorithms (objective video quality metrics) that can provide a measure of the video quality, as perceived by human users. Figure 1.2 shows how the objective and subjective scores are compared to validate the objective quality assessment algorithms. If the

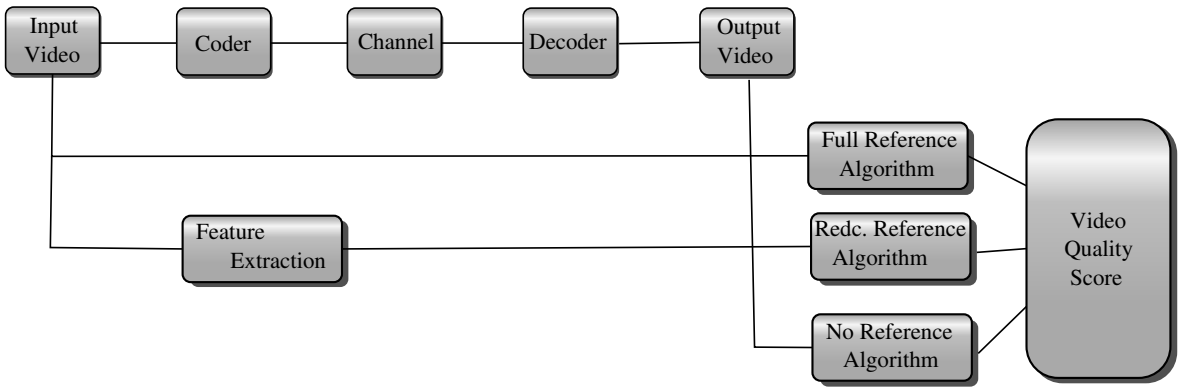


Figure 1.1: Illustration of the 3 types of objective video quality assessment methods: Full Reference (FR), Reduced Reference (RR) and No-Reference (NR).

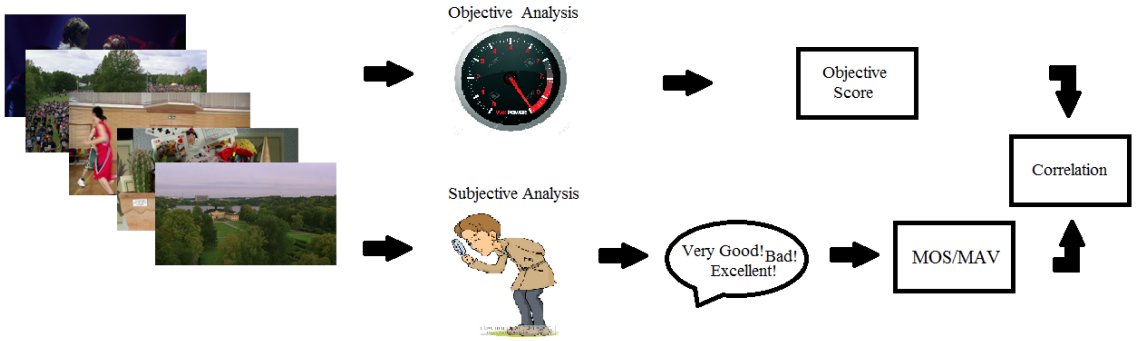


Figure 1.2: Illustration of how the correlation between objective and subjective scores is performed.

correlation index between objective and subjective is close to '1' (or '-1'), the objective quality assessment algorithm is considered to have a good performance.

Video quality is still far from being a mature research topic and limited success has been reported from evaluations of quality models, which are commonly tested under limited conditions with a small diversity of distortions and content [4]. Most of the achievements in the area of video quality have been in the development of full-reference video quality metrics that evaluate annoyance caused by compression artifacts [5,6]. Unfortunately, as mentioned earlier, FR metrics have limited applications and cannot be used in most real-time video transmission applications, like for example, broadcasting and video streaming. In such cases, the undistorted signal (reference) is not available or not accessible at the receiver side and, thereby, requiring even a small portion of it becomes a serious impediment. But, although human observers can usually assess the quality of a video without using the reference, designing a NR (blind) video quality assessment method is a difficult task. Considering the difficulties faced by the FR metrics [7], this is no surprise. A common approach taken by most NR metrics is to try to estimate the strength of the most relevant impairments (e.g, blockiness, blurriness, noise, and ringing) [8].

Among the most important objective video quality metrics, it is worth mentioning some FR, RR and NR works. As an example of an FR metric, we can cite the seminal work of Wang *et al.* [5]

who proposed a quality measure that is based on the picture structural similarity (SSIM). The work of Sarnoff [9] predicts image quality by taking into account a measure of the just noticeable differences (JND) between original and distorted images. As an example of a RR metric, we can cite the metric proposed by Gunawan and Ghanbari [10] that is based on a discriminate harmonic measure of the gain/loss information. Kanumuri *et al.* [11] proposed a RR method that estimates how packet-loss artifacts affect video quality. Finally, the NR metric proposed by Mittal *et al.* [12] builds a 'quality aware' collection of statistical features that is based on a simple space domain natural scene statistic model. The NR metric proposed by Moorthy *et al.* [13] is based on the hypothesis that statistical properties of natural images are altered when they are distorted, making them look unnatural.

1.1 PROBLEM STATEMENT

The Internatinal Telecommunication Union defines Quality of Service (QoS) as a set of charateristics of a telecommunications service that targets user satisfaction [14, 15]. QoS measures are basically physical measurements of the telecommunication services and structure that estimate the system performance. Commonly used QoS measures are jitter, packet loss rate, delay, and bandwith. On the other hand, Quality of Experience (QoE) measures take into account the user satisfaction with the received audio-visual content [15]. More specifically, it measures the overall experience of the final user, taking into account the way humans consume audio-visual contents. It is worth mentioning that QoE is affected, not only by the quality of received signal, but also by the content, physical environment, display device, reproduction playout, and the user expectation.

In subjective experiments, participants usually evaluate video quality by giving a numerical score or an adjective (excellent, good, fair, poor or bad) [16] to describe the quality of the image. It is known that to estimate quality human observers take into account factors like color brightness, light intensity, contrast, sharpness, and the absence/presence of distortions [15]. According to Baraković *et al.* [17], QoE evaluation is affected by the following aspects: (1) technological performance, (2) usability, (3) subjective aspects, (4) expectation, and (5) context. Thus, QoE encompasses more than a score, it represents a numbers of aspects that are considered important by users when watching audio–visual content. Figure 1.3 depicts a possible relation between QoS and QoE [15]. This graph shows that the relation between QoS and QoE is not well defined because QoE depends on several subjective factors. For example, if a video has a very poor quality, viewers might simply not watch it. Therefore, only QoE can be used to quantify acceptability.

As expected, quality metrics play an important role in communications quality control systems. Although there are several subjective factors that influence QoE, the quality of the received content has a direct effect on the user satisfaction and, consequently, on the acceptability of the service [18]. Although there has been a lot of advances in the development of FR image quality assessment methods, the design of a NR metric is still a big challenge. In order to be used in real-time applications, NR methods need to be fast enough to deal with hundreds and thousands of

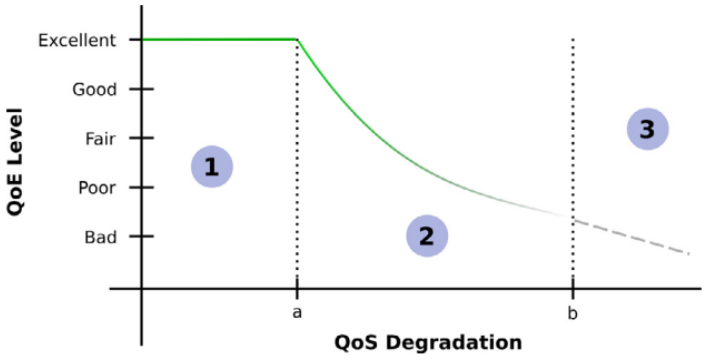


Figure 1.3: Relation between QoS and QoE measures [Extracted from [15]].

frames per second (fps), i.e. the algorithm cannot be complex. Also, one particular scenario that still demands a lot of work is the digital transmission, which includes the internet based video transmission. Up to date, there is no NR video quality metric that can estimate the quality of streamed video in *real-time*.

1.2 PROPOSED APROACH

In the literature, among the quality metrics available for digital video transmission applications, we can cite the work of Garcia *et al.* [19] who implemented an audiovisual metric targeted at network quality monitoring. Farias *et al.* [8] proposed a no-reference metric based on a combination of blockiness, blurriness, and noisiness artifact measurements. Winkler [2] proposed a no-reference hybrid video quality metric that combines network status information and data extracted from the video bitstream. Babu *et al.* [20] studied the effect of block-edge and packet-loss impairments in video streaming applications. So, very few pixel-based NR quality metrics are able to measure degradations introduced by digital transmission, like packet-loss and jerkiness [21].

In this work, we propose a hybrid NR video quality metric that measures three different types of digital transmission degradations: packet-loss, blockiness, and bluriness. The main goal of this work is to develop a no-reference video quality metric for digital transmission applications. The proposed approach is based on the extraction of features that characterize these three different types degradations. A machine learning algorithm is used to combine the extracted features and obtain an estimate for the video quality. The proposed metric is blind, fast, and reliable enough to be used in real-time scenarios.

To validate our approach, the proposed metric is tested using several video quality databases. Since packet-loss is one the of most relevant artifacts in digital transmission, we first design a blind packet-loss metric that is able to detect and estimate the strength of this type of artifact. Then we used the features used in the packet-loss metric, along with features used in blurriness

and blockiness metrics, to design the proposed metric.

1.3 ORGANIZATION

This dissertation is divided as follows. In Chapter 2, we give the background to the material presented in this work, describing basic aspects of video transmission and compression, common video degradations, popular video quality metrics, statistical measures and common video quality databases. In Chapter 3, we propose the proposed methodology. In Chapter 4, we present our results. Finally, in Chapter 5, we present future works and conclusions.

2 DIGITAL VIDEO BACKGROUND

In this chapter, we present the basic background needed for this work, including a very brief description of video compression algorithms, a description of the most common video degradations, a review of a set of popular video quality metrics, and a review of the machine learning algorithms and statistical measures used in this work.

2.1 VIDEO COMPRESSION

Compression refers to the process of reducing the amount of data required to represent a given quantity of information [22]. There are two types of compression: lossless and lossy. In lossless compression, data is perfectly reconstructed without any loss when compared to the original data. Huffman [22], Gollomb [22], Arithmetic [22], LZW [22] are some examples of these types of compression algorithms. In lossy compression, the algorithm discards nonessential information that is considered less perceptually relevant, according to human visual system models [22]. This allows for a much higher compression rate. Examples of lossy compression algorithms are the JPEG family of image compression algorithms and the MPEG and the ITU families of video compression algorithms [22]. For instance, H.264 (joint MPEG/ITU algorithm) is used in the Brazilian Digital TV system.

Given that our work focuses on video, we briefly describe the family of MPEG/ITU compression algorithms. MPEG uses the same principles of the JPEG algorithms, which is a DCT block-based lossy compression algorithm [22]. Besides reducing spatial redundancy like the image compression algorithms, video compression algorithms also reduce the temporal redundancy between video frames. With this goal, most video compression algorithms use motion estimation and motion compensation algorithms. In the family MPEG compression algorithms, the video frames are split into 8x8, 16x16 or 32x32 pixels depending on the type codec you use. MPEG-2 frames are divided in macroblocks (MB) of 16x16 pixels. To estimate motion, each MB in the current frame is compared to a set of MB in the previous frame, as shown in Figure 2.1. The similarity between these MBs is calculated by taking the sum of absolute differences (SAD) measure. The most similar MB in the previous frame is chosen as a 'prediction' of the current MB. The algorithm encodes the difference between current and predicted MB position and the position to the predicted MB.

Two types of frame coding are used in MPEG: intra and inter frame coding. If intra frame coding is used, the coding is performed independently of previous (or subsequent) frames. On the other hand, if inter frame coding is used, MB in the current frame are compared to MB in the previous (or subsequent) frames and only the prediction error and the MB position is coded. Frames in MPEG are split into Group of Pictures (GOP), which contains an I (intra) coded frame

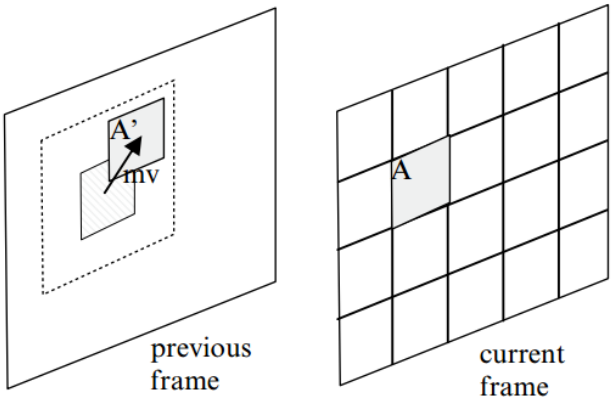


Figure 2.1: Interframe Motion Compensation (Extracted from [23]).

Figure 2.2: Example of a video frame severely affected by packet-loss artifacts.

and a set of P (previous) and B (bidirectional) coded frames. One parameter of the algorithm is the number of P and B frames in a GOP.

2.2 VIDEO DEGRADATIONS

In this section, we briefly describe the most common degradations present in digital video transmission scenarios: packet-loss, blockiness, and bluriness.

2.2.1 Packet-Loss Artifacts

Packet-loss is an artifact that is generated when packets are lost during the transmission process. Packets may be lost during severe channel transmission conditions or severe traffic conges-

tions. You can lose MPEG slices, IP packets or TS packets. When packets are lost, information corresponding to areas in the video frames are lost. Compression algorithms generally use error concealment algorithms to mask these losses. But, if the lost packets belong to I and P frames, the degradations may propagate for several frames. Therefore, the impact of packet losses on the video quality depends on their corresponding temporal and spatial position in the frame and the relevance of their content.

Figure 2.2 illustrates the visual effect of a packet-loss in a frame. Notice that, the blocks affected by the packet-loss artifact can be easily perceived. This happens because, although the decoder tries to predict the missing information, it fails to recover all original information. In this figure, we can notice how the visibility of the artifact is affected by content. For example, the robot's head, arms, body, and background are severely affected by the packet-loss artifact. When viewers watch this video, they perceive it as having a poor quality because of the presence of these highly annoying artifacts.

It is worth mentioning that packet-loss artifacts do not occur uniformly over the frame. As mentioned earlier, their behavior depends on the digital transmission conditions. As a consequence, distortions can affect any part of the picture frame and, as shown in Figure 2.2, these distortions are not correlated to one another. In other words, not all frames may be affected by packet-loss artifacts, which are also generally not uniformly distributed in time and space.

2.2.2 Blockiness Artifacts

Blockiness is an artifact that is generated as a result of the compression process. As mentioned earlier, compression algorithms split frames into blocks. Each block is coded independently, i.e. different compression parameters may be used for different (sometimes neighboring) blocks. The blockiness effect is created when differences between neighboring blocks is visible, what is more frequent in high compression rates.

Figure 2.3 depicts the effect of a blockiness artifact in a video frame . In this particular frame, the blockiness artifacts are spread all over the picture frame. It is important to mention that the blockiness artifact affects the picture frame uniformly. However, depending on the content, blockiness may be more or less visible. For instance, it is easier to see the artifact in the areas showing 'a man riding a bike' and 'a woman wearing a pink t-shirt and black pants'. Nevertheless, the artifact is less visible in the buildings in the background.

2.2.3 Blurriness Artifacts

Blurriness is an artifact that is generated as a consequence of discarding high frequencies, causing the image to loose sharpness or have softer edges. Figure 2.4 shows the blurriness effect in a frame. Although there are seven people in the picture, the viewer cannot clearly identify all of them because of how blurred the picture looks. It is also possible to see the effect on the grass,

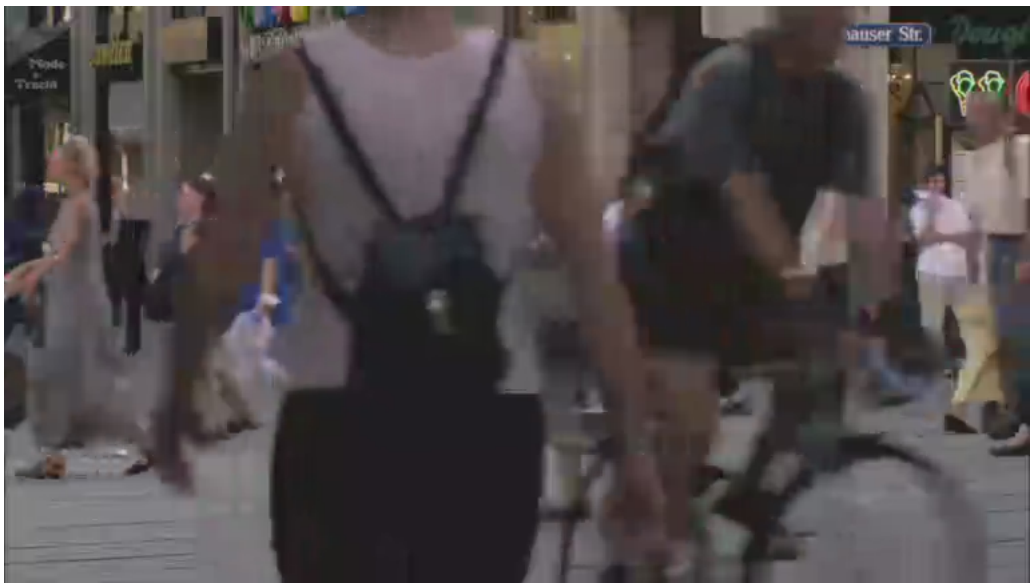


Figure 2.3: Example of a video frame containing blockiness artifacts.

Figure 2.4: Example of a video frame containing blurriness artifacts.

trees and river.

2.3 OBJECTIVE QUALITY ASSESSMENT METHODS

In this section, we describe a set of quality metrics that are currently used in image quality. The set includes FR and NR metrics that are considered relevant to this work.

2.3.1 Data Fidelity Metrics

Data metrics are metrics which simply compare the pixels of an image or a video, without taking into account their content and their relation to other pixels. One of the most famous FR data metrics is the mean square error (MSE). MSE has been widely used in signal processing applications due to its simplicity and physical meaning. MSE is calculated using the following equation:

$$MSE(I_o, I_d) = \frac{1}{MN} \sum_{n=1}^{N} \sum_{m=1}^{M} (I_o(n,m) - I_d(n,m))^2 \tag{2.1}$$

where M and N are the total number of rows and columns of the image. Notice that MSE is fast, easy to understand and implement. As any quantitative full reference quality metric, MSE requires the original (reference), I_o, and the distorted (test),I_d , images. So, MSE can be also classified as a signal fidelity measure between "distorted" and original images.

Wang Z. and Bovik A. analyzed [24] the advantages and drawbacks of MSE. They showed MSE has the following advantages:

(a) It is simple;

(b) All l_p norms are valid: nonnegativity, identity, symmetry and triangular inequality;

(c) It has a clear physical meaning;

(d) It is an excellent metric in the context of optimization, statistics and, estimation;

MSE also works well for the same type of content and the same type of degradations. Despite these interesting properties, MSE does not have a good correlation with the quality as perceived by human observers [24].

Another FR data metric that is commonly used in image processing is the Peak signal-to-noise ratio (PSNR), which is calculated with the following equation:

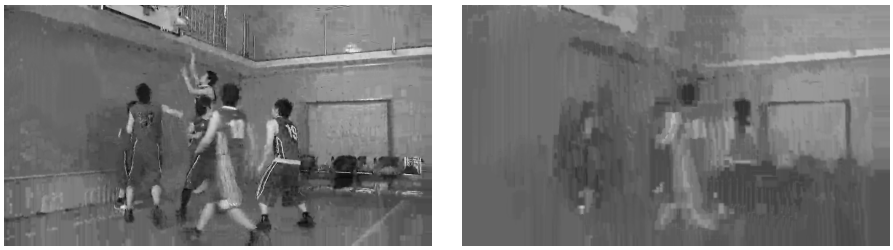
$$PSNR(I_o, I_d) = 10 \log_{10} \frac{L^2}{MSE(I_o, I_d)} \tag{2.2}$$

where L is the maximum possible value of the pixel, which is generally 255 for an eight-bit image.

One of the major reasons why MSE and PSNR (or any other data metric) do not perform as desired is because they do not incorporate properties of the human visual system (HVS) in their computation. Measurements produced by data metrics are simply based on a pixel to pixel (or bit to bit) comparison of the data, without considering what is the content and the relationships among the pixels in an image (or frames). For example, MSE and PSNR do not consider how spatial and frequency components are perceived by human observers [24].

Figure 2.5 shows a comparison of four images with different PSNR and MSE values. Notice that, the higher the PSNR and MSE values, the higher the image quality. Therefore, PSNR and

(a) PSNR=35 dB, MSE=20.56 (b) PSNR=30 dB, MSE=65.02



(c) PSNR=25 dB, MSE=205.63 (d) PSNR=20 dB, MSE=650.25

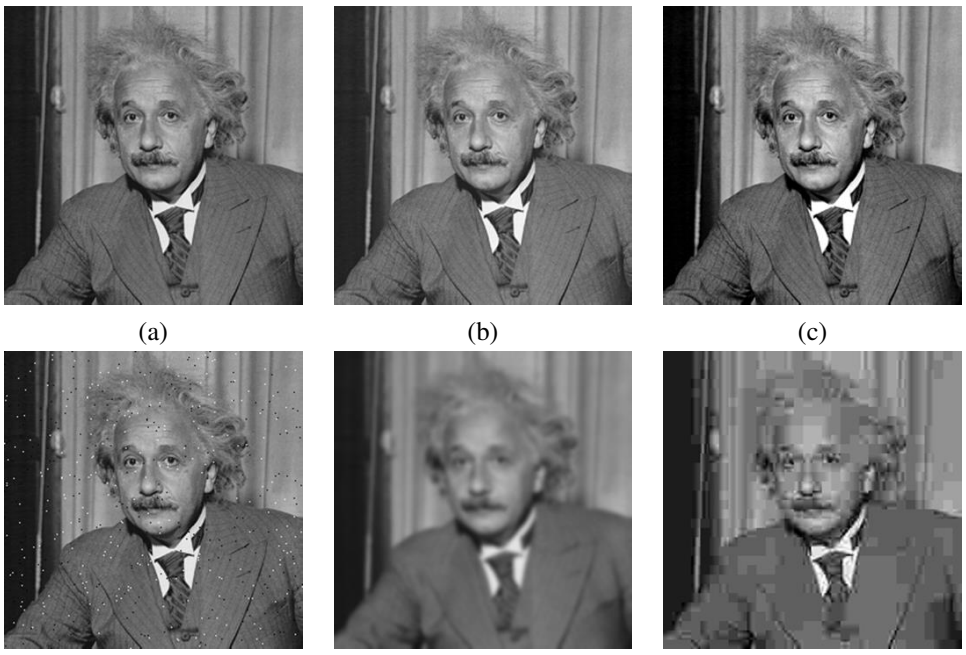
Figure 2.5: Comparison between different PSNR.

MSE seem to be working as a quality measure in these cases. On the other hand, Figure 2.6 depicts 5 distorted images that have approximately the same MSE=144 (PSNR=26.55 dB). Notice that the perceived qualities of these images are quite different. For example, the images in Figures 2.6.(b) and 2.6.(c) have good quality levels, but the images in Figures 2.6.(d), 2.6.(e), and 2.6.(f) have poor quality levels. So, we can conclude that MSE and PSNR are not good metrics to estimate image quality in these cases, for which different distortions or image processing algorithms are being compared.

2.3.2 Pixel Based Image Quality Metrics

Pixel-based metrics are metrics which estimate quality by analyzing the pixels of the image or the video (decoded). These metrics generally take into account image characteristics and human visual models. Two basic approaches are generally used in their design: a vision modeling approach and an engineering approach. The vision modeling approach explicitly incorporates aspects of Human Vision System (HVS) into the algorithm. Nevertheless, it is important to point out that knowledge about how the HVS works is still incomplete. Therefore, although image processing and computer vision algorithms use these models to improve their efficiency, HVS models are somewhat imprecise.

The engineering approach is based on the extraction of specific features of the visual signal, like for example specific spatial and temporal frequencies, statistical measures, sharpness and contrast measures, etc. A very popular type of features are the features that indicate the presence of specific artifacts, like blockiness, ringing, and blurriness. This approach has been widely used because of its simplicity. In this section, we describe three popular pixel-based metrics.



(a) (b) (c)

(d) (e) (f)

Figure 2.6: Comparison between the different pictures with the same MSE. (a) Original, MSE=0,SSIM=1 (b) Unsharped, MSE=144, PSNR=26.55 dB, SSIM=0.988 (c) MSE=144, PSNR=26.55 dB, SSIM=0.913 (d)MSE=144, PSNR=26.55 dB, SSIM=0.840 (e) MSE=144, PSNR=26.55 dB, SSIM=0.694 (f) MSE=144, PSNR=26.55 dB, SSIM=0.662

2.3.2.1 SSIM

The Structural Similarity (SSIM) is a full reference image quality assessment method developed by Wang [25]. SSIM compares three features of the reference and distorted image. These features are luminance, contrast and structure. It performs its analysis on the luminance component of the image. To compute the predicted score, the image is first divided in 8x8 blocks. For each block the luminance, contrast and structure comparison measurements are performed.

The luminance comparison is calculated by the following equation:

$$l(I_o, I_d) = \frac{2\mu_o\mu_d + C}{\mu_o^2 + \mu_d^2 + C}, \tag{2.3}$$

where μ_o and μ_d are the mean intensity of the original image block and the distorted (test) image block, respectively. And C is a small constant necessary to avoid instability.

The Contrast Comparison is calculated by the following equation:

$$c(I_o, I_d) = \frac{2\sigma_o\sigma_d + C_2}{\sigma_o^2 + \sigma_d^2 + C_2}, \tag{2.4}$$

where σ_o and σ_d are the standard deviation intensity of the original and distorted (test) image block, respectively. And C_2 is a small constant necessary to avoid instability.

The Structure Comparison is calculated by the following equation:

$$s(I_o, I_d) = \frac{\sigma_{od} + C_3}{\sigma_o \sigma_d + C_3}, \tag{2.5}$$

where σ_o and σ_d are the standard deviation intensity of the original and distorted block, σ_{od} is the covariance between the original image block and the distorted (test) image block, and C_3 is a small constant necessary to avoid instability.

To combine these comparisons into a single SSIM map, we use the following equation:

$$SSIM(I_o, I_d) = l(I_o, I_d)^\alpha \cdot c(I_o, I_d)^\beta \cdot l(I_o, I_d)^\gamma, \tag{2.6}$$

where, to simplify, α = β = γ = 1. The average value of the SSIM map is the final score. To use SSIM as a video quality method, we calculate the SSIM for each video frame and average these values to obtain the video quality score.

As mentioned earlier, PSNR and MSE are not good quality metrics for the images in Figure 2.6. Nevertheless, this figure also shows the SSIM values for each image and we can notice that SSIM is able to differentiate the different levels of quality of the images. For example, the good quality images in Figures 2.6.(b)and 2.6.(c) obtain high SSIM values, while the lower quality images in Figure 2.6.(d), 2.6.(e), and 2.6.(f) smaller SSIM values.

2.3.2.2 Blockiness NR Metric

Wang [26] proposed a NR metric to estimate the strength of blockiness artifacts. Figure 2.7 shows a block diagram of this algorithm. The algorithm takes the differences between consecutive columns, as shown in the following equation:

$$d_h(m, n) = I_d(m, n + 1) - I_d(m, n) \tag{2.7}$$

where I_d is the luminance component of he video frame, M is the number of rows and N number of columns.

The horizontal blockiness measure ($B_h(k)$) is computed using the following equation:

$$B_h(k) = \frac{1}{M(\lfloor N/8 \rfloor - 1)} \sum_{i=1}^{M} \sum_{j=1}^{\lfloor N/8 \rfloor - 1} |d_h(i, 8j, k)| \tag{2.8}$$

where k is the frame number, which ranges from 1 to the NF. The vertical measure ($B_v(k)$) is obtained in a similar way.

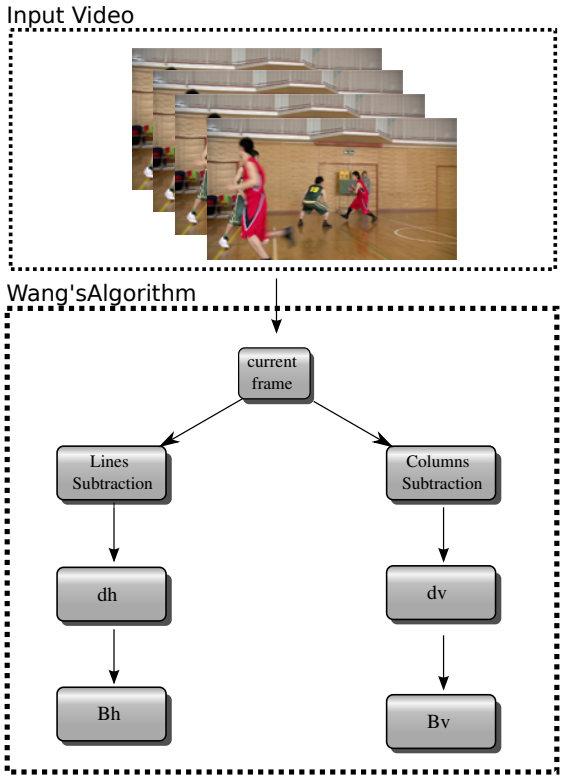


Figure 2.7: Wang's Blockiness NR Metric.

The final horizontal and vertical blockiness scores are given by:

$$WH = \sum_{k=1}^{NF} Bh(k) \tag{2.9}$$

and

$$WV = \sum_{k=1}^{NF} Bv(k) \tag{2.10}$$

where NF is the number of frames and k the frame number. The features WH and WV will be used to compose the hybrid metric.

2.3.2.3 Bluriness NR Metric

Crété-Roffet [27] proposed a NR metric that quantifies the strength of bluriness. Figure 2.8 shows a block diagram of this algorithm. It compares the differences between neighboring pixels, before and after the image is low-passed filtered and the luminance component (Y) is filtered in vertical and horizontal directions separately. The filters are described by the following equations:

$$h_h = h_v^T = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}. \tag{2.11}$$

Let's denote $I(x, y, k)$ as the k-th frame, $BL_h(x, y, k)$ as the blur horizontal image and $BL_v(x, y, k)$

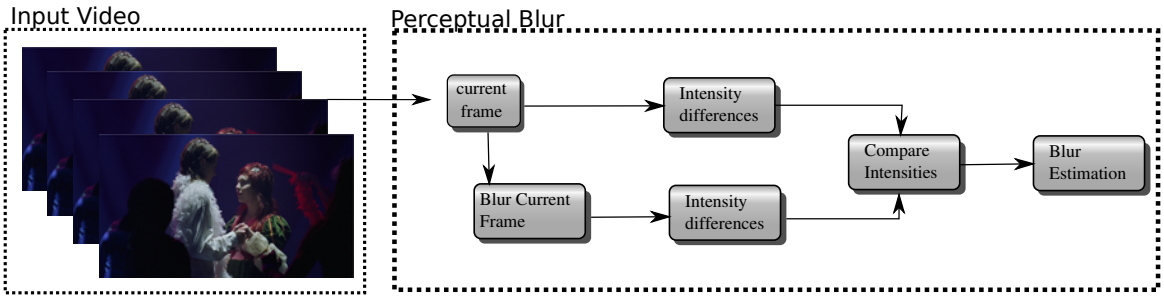


Figure 2.8: Perceptual Blur NR Metric.

as the blur vertical image corresponding to this frame. Next, we calculate the horizontal and vertical intensity differences of the frame $I(x, y, k)$:

$$ID_H(k) = \sum_{i=0}^{M-1} \sum_{j=1}^{N-1} |I(i, j, k) - I(i, j-1, k)| \tag{2.12}$$

and

$$ID_V(k) = \sum_{i=1}^{M-1} \sum_{j=0}^{N-1} |I(i, j, k) - I(i-1, j, k)|. \tag{2.13}$$

where $I(x, y, k)$ is the luminance component of the frame, M is the number of rows, N number of columns and k is the frame index, ranging from 1 to NF. At the same time, we calculate maximum difference between neighboring pixels in $I(x, y, k)$ and $BL_h(x, y, k)$ or $BL_v(x, y, k)$, which are given by the following equations:

$$MD_H(k) = \sum_{i=0}^{M-1} \sum_{j=1}^{N-1} \max(0, |I(i,j,k) - I(i,j-1,k)| - |BL_h(i,j,k) - BL_h(i,j-1,k)|) \tag{2.14}$$

and

$$MD_V(k) = \sum_{i=1}^{M-1} \sum_{j=0}^{N-1} \max(0, |I(i,j,k) - I(i-1,j,k)| - |BL_v(i,j,k) - BL_v(i-1,j,k)|) \tag{2.15}$$

Finally, bluriness is estimated using the following ratio:

$$blur(k) = \max\left(\frac{ID_H(k) - MD_H(k)}{ID_H(k)} \quad \frac{ID_V(k) - MD_V(k)}{ID_V(k)}\right) \tag{2.16}$$

Notice that, the higher the value of this ratio, the higher the strength of the blurriness. For the complete video, the blur measure is obtained by taking the average of the $blur$ estimate, given in equation 2.16, for all NF frames.

The features ID_V, ID_H, MD_H and MD_V are used to compose the hybrid metric.

2.3.2.4 Correlation Based Packet-loss Metric

Vlachos [42] proposed an NR metric to estimate the strength of blockiness artifacts in video signals. The algorithm compares the cross-correlation of pixels inside (intra) and outside (inter) the borders of the coding blocking structure of a video frame $I(i, j, k)$, considering blocks of 8×8 and a downsampled version of the frame in vertical and horizontal directions.

In a previous work Farias [8, 43] modified Vlachos' algorithm [42]. Farias proposed to split the down-sampling process into separate vertical and horizontal directions. As a consequence, a vertical downsampled image (SV) and a horizontal downsampled image (SH) are generated using the following equations:

$$SH_m = \{Y(i, j) : m = i \mod 8\}. \tag{2.17}$$

$$SV_n = \{Y(i, j) : n = j \mod 8\}. \tag{2.18}$$

where (i, j) are the horizontal and vertical co-ordinates and $\mod$ is the module operation. This way, SV_n and SH_m contain subsets of pixels with coordinates congruent to 8, either horizontally or vertically, respectively. The subscripts m and n can be viewed as the corresponding horizontal and vertical phases, respectively.

Figures 2.9 (a) and (b) display the sampling structures used by Farias' in the horizontal (SH_m) and vertical (SV_n) directions, respectively. The image shows a 16×16 area of the frame, containing four 8×8 blocks. Six sub-images are generated by downsampling pixels located at the positions indicated by the six different symbols. Therefore, different symbols generate different sub-images. The set of inter-block pixels in the vertical direction corresponds to the sub-images SV_0 and SV_7 (Figure 2.9 (b)), while the set of inter-block pixels in the horizontal direction corresponds to the sub-images SH_1 and SH_7 (Figure 2.9(a)). The set of intra-block pixels in the vertical direction corresponds to the sub-images SV_0 and SV_1 (Figure 2.9 (b)), while the set of intra-block pixels in the horizontal direction corresponds to the sub-images SH_1 and SH_3 (Fig. 2.9 (a)).

Given that interlaced videos were used by Farias, the symbols in the horizontal downsampling structure (see Figure 2.9 (a)) are 2 pixels apart, instead of only one pixel like in the vertical downsampling structure (see Figure 2.9 (b)). For progressive videos, the symbols should be one pixel apart for both directions. Figure 2.10 displays how the sub-image SV_0 is obtained. In this example, the original frame has 1280×720 pixels and the vertically-downsampled sub-image has 160×720 pixels.

The cross-correlation between two frames, I_1 and I_2, is given by the following expression:

$$C_{I_1,I_2}(i, j) = F^{-1}\left(\frac{F^*(I_1(i, j)) \cdot F(I_2(i, j))}{|F^*(I_1(i, j))F(I_2(i, j))|}\right), \tag{2.19}$$

where F and F^{-1} denote the forward and inverse two dimensional discrete Fourier transform, re-

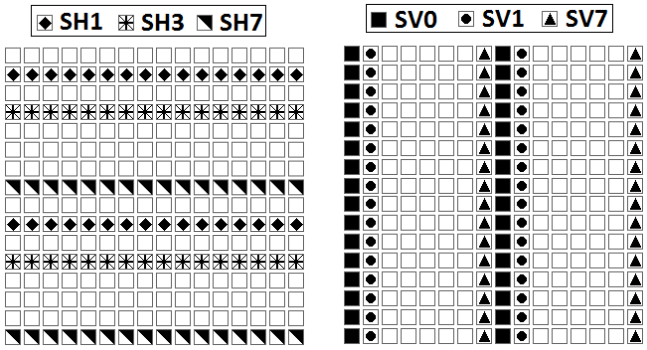


Figure 2.9: Frame downsampling structure for: (a) horizontal and (b) vertical directions.

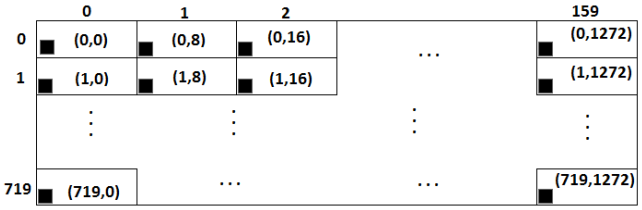


Figure 2.10: Illustration of vertical downsampling process used to obtain the sub-image SV_0.

spectively, and $*$ denotes the complex conjugate. The magnitude of the highest peak is a measure of the correlation between I_1 and I_2. But, before the maximum is calculated, the array elements is filtered using a Hamming window, what forces the elements to a constant value around the borders.

To estimate the blockiness signal strength, Farias measured the correlation between the intra- and inter-block sub-images in both directions. For the vertical direction, the correlation was calculated using the following equations:

$$CV_{intra}(k) = \max_{i,j} \{C_{SV_0,SV_1}(i,j,k)\}\,, \tag{2.20}$$

$$CV_{inter}(k) = \max_{i,j} \{C_{SV_7,SV_0}(i,j,k)\}\,. \tag{2.21}$$

The horizontal correlations, $CH_{inter}(k)$ and $CH_{intra}(k)$, were obtained in a similar way:

$$CH_{intra}(k) = \max_{i,j} \{C_{SH_1,SH_3}(i,j,k)\}\,, \tag{2.22}$$

$$CH_{inter}(k) = \max_{i,j} \{C_{SH_7,SH_1}(i,j,k)\}\,. \tag{2.23}$$

Then, the blockiness measure for one frame was given by:

$$S_{bloc}(k) = \frac{CV_{intra}(k) + CH_{intra}(k)}{CV_{inter}(k) + CorrH_{inter}(k)}. \tag{2.24}$$

For frames with no blockiness, the value of $CV_{intra}(k)$ was close to $CV_{inter}(k)$ and $CH_{intra}(k)$

was close to $CH_{inter}(k)$. As blockiness was introduced, the values of $CV_{inter}(k)$ and $CH_{inter}(k)$ became smaller and, consequently, the value of the blockiness metric increased.

Finally, the blockiness measure for the set of all frames was obtained by averaging the measures over all frames:

$$Bloc = \frac{1}{NF} \sum_{k=0}^{NF} S_{bloc}(k), \tag{2.25}$$

where k refers to the frame number and NF is the total number of frames.

2.3.3 Hybrid Metrics

A hybrid metric is a metric that combines different types of pixel-based metrics to obtain a single quality estimate [2]. For example, we may combine blockiness, ringing, and blurriness metrics to obtain a quality metric targeted at compression applications. In this work, we propose to use this approach to combine features corresponding to degradations that are common in digital transmission scenarios. In particular, we consider blockiness, blurriness, and packet-loss artifacts. The features are combined using a Support Vector Regression (SVR) technique, which is detailed in the next section. The metric is trained on a set of video quality databases which are described in Section 2.4. More details of the proposed methodology will be given in the next chapter.

2.4 VIDEO QUALITY DATABASES

Objective video quality assessment methods are costumarily validated using anotated video quality databases. These databases consist of a set of videos with a diverse content, which are processed with different Hypothetical Reference Circuits (HRC). Different HRCs generate test videos with different types of artifacts at different levels of annoyance. For each video in the database, there is a corresponding mean observer score (MOS).

In this work, five video quality databases are used: Varium (Sets 1-3) [32–34], Roma [35–37], Live [38, 39], CSIQ [40] and IVPL [41]. Table 2.1 shows a summary of characteristics of all databases.

Table 2.1: Video Quality Database parameters.

Database	No.videos	Format	Spatial Res.	Temporal Res.	Time Dur.	Distortions
Roma	184	4:2:0	704x576 720x576	25,30	7-9	Packet loss rate, jitter, delay, and throughput
Live	150	4:2:0	768x432	25,50	10, 8.68	H264,,MPEG2,IP error and wireless networks
CSIQ	216	4:2:0	832x480	24,25,30, 50,60	10	H.264, HEVC/H.265, Wavelet-based compression wireless, transmission loss and AWGN
IPVL	128	4:2:0	1920x1088	25	8.96,10,11.2	H.264, MPEG2, Dirac coding, IP error
Varium						
Set 1	84	4:2:0	1280x720	50	10	Packet-Loss
Set 2	119	4:2:0	1280x720	50	10	Blockiness and Blurriness
Set 3	140	4:2:0	1280x720	50	10	Packet-loss, Blockiness and Blurriness

2.4.1 Varium

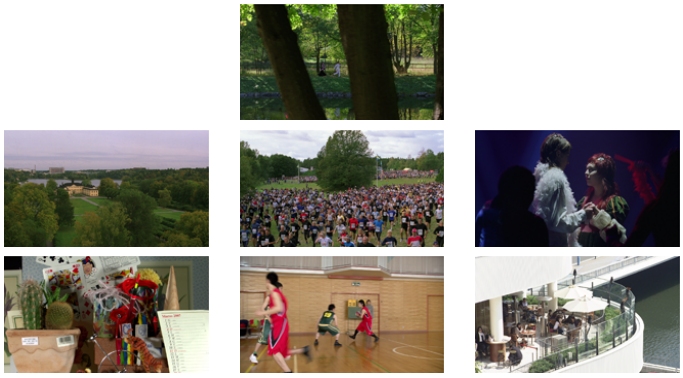


Figure 2.11: Sample frames of originals of the Varium database.

The Visual Artifacts Interference Understanding and Modeling (Varium) is a database that contains the results of six experiments [32–34]. The database is divided into three sets of test sequences. Set 1 contains sequences only with packet-loss artifacts at different strengths (packet-loss rates: 0.7%, 2.6%, 4.3%, and 8.1%) and different durations (4, 8, and 12 frames). Set 2 contains videos with blockiness and bluriness artifacts. Finally, Set 3 contains videos with the three types of artifacts.

For each set, two types of experiments were performed: annoyance and strength experiments. In the annoyance experiments, participants were asked to rate the annoyance of the degradations present in the test videos. In the strength experiments, participants are asked to rate the perceptual strength of each type of artifact in the videos. Therefore, for each set in Varium, we have mean annoyance (MAV) and mean strength values (MSV) scores for each test sequences.

Videos in Varium have a spatial resolution of 1280×720 and a temporal resolution of 50 frames per second (fps). They are all ten seconds long and were chosen with the goal of including diversity in content in the test set. Figure 2.11 depicts sample frames of the original videos in the Varium database.

Figure 2.12: Sample frames of originals of the Roma database.

2.4.2 Roma database

ReTRiEVED Video Quality Database (Roma) database [35–37] contains 184 distorted videos, with a color format YVV 4:2:0, spatial resolutions of 704×576 and 720×576, temporal resolution of 25 and 30 frames per second (fps), and durations of 7, 8 or 9 seconds. This database contains up to four types of distortions (HCRs): packet-loss rate, jitter, delay, and throughput. In this work, we only used the test videos with packet-loss artifacts, given that the other distortions are not measurable by pixel-based metrics. Figure 2.12 depicts sample frames of the original videos in the Roma database.

Figure 2.13: Sample frames of originals of the Live database.

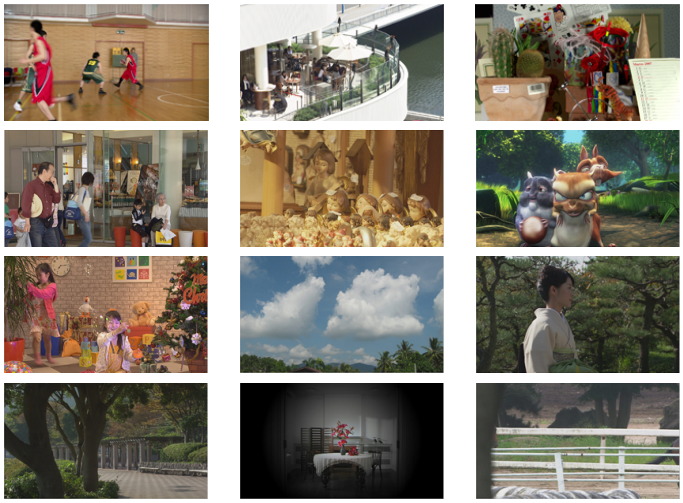


Figure 2.14: Sample frames of originals of the CSIQ database.

2.4.3 Live database

The Live database [38, 39] consists of 150 videos, with YUV format YUV 4:2:0, spatial resolution of 768x432 pixels, temporal resolution of 25 and 50 fps, and durations of 10 and 8.68 seconds long. The database contains videos with up to four types of distortions: H264, MPEG2, IP error and wireless networks. Figure 2.13 depicts sample frames of the original videos.

2.4.4 CSIQ database

The CSIQ database [40] contains 216 distorted videos with a spatial resolution of 832x480 pixels, format YUV 4:2:0, duration of 10 seconds long, temporal resolution of 24, 25, 30, 50 and 60 frames per second. The database contains videos with up to six types of distortions: H.264, HEVC/H.265, MJPEG, Wavelet compression, Wireless and AWGN Noise.

2.4.5 IVPL database

The IPVL database [41] consists of 128 distorted videos, 10 reference videos (originals) with spatial resolution of 1920×1088 progressive and temporal resolution of 25 frames per second (fps). The Distortion types are H.264, MPEG2, Dirac coding, IP error. Figure 2.15 shows the original videos of the IVPL database.

Figure 2.15: Sample frames of originals of the IVPL database.

3 PROPOSED HYBRID NO-REFERENCE VIDEO QUALITY METRIC

In this chapter, we detail the hybrid video quality assessment methodology developed in this work. Figure 3.11 depicts a block diagram of this approach. Notice that the proposed methodology consists of a hybrid approach that combines the features gathered by three individual artifacts metrics. A support vector regression (SVR) algorithm combines these features to obtain an overall quality estimate. Contrary to previous approaches, SVR uses features collected by the artifacts metrics and not the overall objective scores obtained by these metrics. Before describing the hybrid approach, we present a packet-loss artifact metric that is designed as part of this work.

3.1 CORRELATION BASED PACKET-LOSS METRIC - PROPOSED METRIC

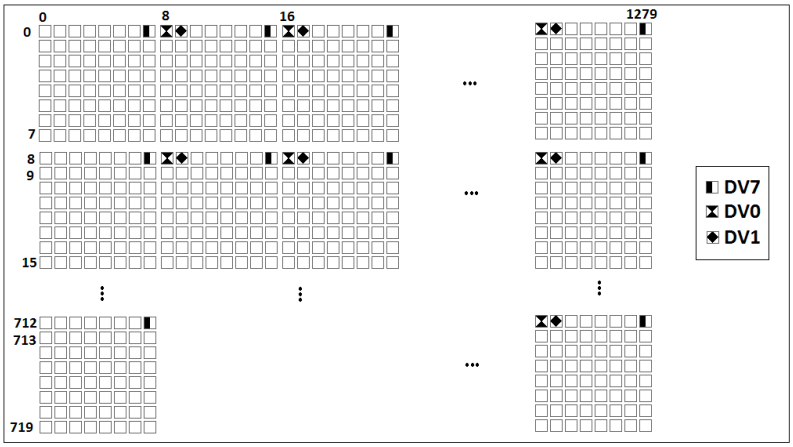
The proposed no-reference blockiness metric is based on the blockiness metric described in the section 2.3.2.4 . To adapt the metric proposed by Farias [8,43] to measure packet-loss (instead of blockiness), we first vary the size of the downsampling structure. Since videos compressed with modern codecs (like H.264 and H.265) use macroblocks of several sizes, we generalize the algorithm proposed by Farias for 8×8, 16×16, and 32×32 block sizes. Figures 3.1 (a) and (b) show the 8×8 vertical and horizontal downsampling frame structures. Again, the dark symbols in the grids correspond to pixels in the resulting downsampled sub-images. The sampling structures for 16×16 and 32×32 are similar. Differently from the algorithm proposed by Farias (see Fig. 2.9), the proposed algorithm simultaneously downsamples the original frame in both directions, reducing the size of the original image in both dimensions.

A total of 6 downsampled images are obtained after the downsampling process, with three sub-images being obtained from the vertical downsampling (DV_7, DV_0, DV_1) and three sub-images from the horizontal downsampling (DH_7, DH_0, and DH_1). Then, we calculate the cross-correlation between two sub-images to obtain the blockiness measure for a single frame (see equations 3.1, 3.2, 3.3, 3.4). More specifically, for the vertical direction, we obtain the inter-block correlation by calculating the correlation between sub-images DV_7 and DV_0 and the intra-block correlation by calculating the correlation between sub-images DV_0 and DV_1:

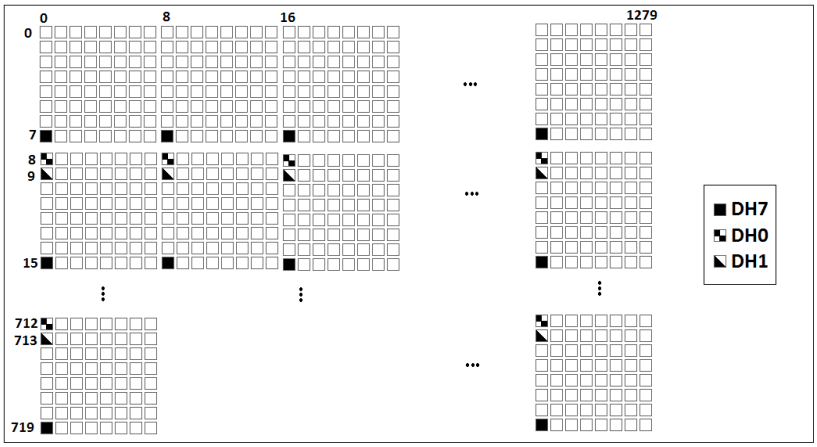
$$CV_{intra,8}(k) = \max_{i,j} \{C_{DV_0,DV_1}(i,j,k)\}\,, \tag{3.1}$$

$$CV_{inter,8}(k) = \max_{i,j} \{C_{DV_7,DV_0}(i,j,k)\}\,. \tag{3.2}$$

Similarly, for the horizontal direction, we obtain the inter-block correlation calculating the correlation between sub-images DH_7 and DH_0 and the intra-block correlation calculating the



(a)



(b)

Figure 3.1: Frame downsampling structure for the proposed packet-loss metric: (a) vertical and (b) horizontal.

correlation between sub-images DH_0 and DH_1:

$$CH_{intra,8}(k) = \max_{i,j} \{C_{DH_0,DH_1}(i,j,k)\}\,, \tag{3.3}$$

$$CH_{inter,8}(k) = \max_{i,j} \{C_{DH_7,DH_0}(i,j,k)\}\,. \tag{3.4}$$

The 8×8 block measure for the k-th frame is given by:

$$S_{bloc,8}(k) = \frac{CV_{intra,8}(k) + CV_{inter,8}(k)}{CH_{intra,8}(k) + CH_{inter,8}(k)} \tag{3.5}$$

Notice that, given that we are assuming the frames are in a progressive format, there is no shift between the pixels. To obtain a measure for the complete video, we average $S_{bloc,8}(k)$ for all frames, obtaining $Bloc_8$.

Next, we use the same algorithm on blocks of size 16×16 ($Bloc_{16}$) and 32×32 ($Bloc_{32}$). The final packet-loss metric value is a composition of the measures for the three block sizes ($Bloc_8$, $Bloc_{16}$, and $Bloc_{32}$), which is obtained using a support vector regression (SVR) [28–31] tech-

nique. We choose to use SVR because similar machine learning-based approaches have been used with success to model complex non-linear perceptual processes related to artifact annoyance [44].

The correlation based packet-loss metric influences two characteristics in the Intensity Difference Packet-loss Metric. First, blocks cannot be considered uniformly in a frame, i.e. all blocks are taken into account to quantify the strength of the artifact. Instead, a detection algorithm needs to be implemented to identify frame areas where blocks were affected by packet-loss. Second, three types of block size (8x8, 16x16, 32x32) are used, it happens because the correlation values increases when these different blocks are considered. Therefore, these ideas were used in the Intensity Difference metric.

3.2 INTENSITY DIFFERENCE PACKET-LOSS METRIC

The intensity difference packet loss metric has two stages: detection and measurement. The detection stage has the goal of identifying which areas of the video contain packet-loss artifacts. The second stage has the goal of measuring the number of packet-losses in the video and estimating their perceptual strength.

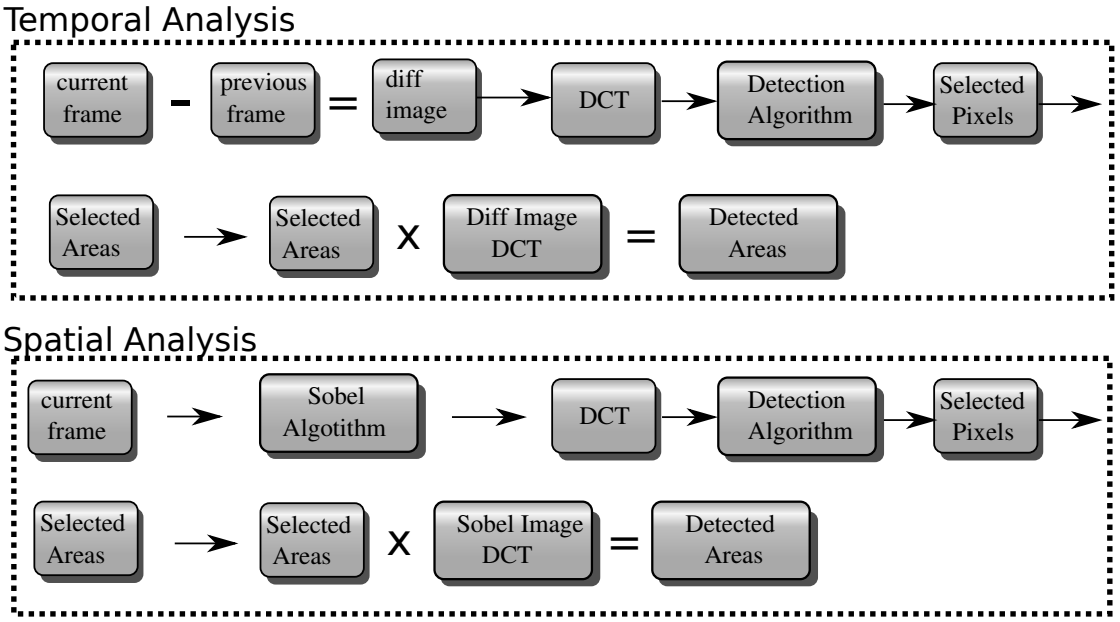


Figure 3.2: Block Diagram of the Intensity Difference Packet Loss Metric.

Figure 3.2 shows a block diagram of the packet-loss detection stage. First, the difference between the current and previous frame is calculated to remove any possible camera movements and noise. After that, the difference map is split into 8×8 blocks and the DCT of each block is calculated. It is worth pointing out that frames are divided in blocks 8×8 because the reference detection algorithm [45] is performed this way. Since packet-loss artifacts are characterized by the presence of strong edges, the detection stage aims to identify regions that contain high edges.

More specifically, the algorithm uses the DC and AC DCT coefficients to detect these strong edges. Bhattacharyya *et al.* [45] showed that it is possible to measure strong edges in the DCT domain. They compared the DC coefficient of the luminance component of each block with the DC coefficients of its immediate spatial neighbors. If the module of the energy difference was higher than 150, their algorithm assumed that an edge was present in this location. AC components were also taken into account. The sum of the module of the first five AC components had to be higher than 50 for the area to be marked as an edge.

We propose a metric that is a modification of Bhattacharyya's metric with AC and DC coefficients conditions. First, instead of comparing the immediate spatial neighbors, we compare the current blocks with the blocks immediately below. The module of the DC energy difference has to be higher than 50, given by the following equation:

$$|DC_{current} - DC_{below}| > 50 \tag{3.6}$$

Second, the procedure for taking the AC components is the same as Bhattacharyya's, i.e. the sum of module of the first five AC components has to be higher than 50. The AC threshold is taken of his work [45]. Three images are generated as a result of this process: a image of selected pixels, an image of selected areas and an image of detected areas. Figures 3.3, 3.4, 3.5 show the three images obtained for a video frame from the Varium database.

Figure 3.3 (selected pixels) shows the selected high edges points that identify the packet-loss artifacts. Each selected pixel is expanded into a 64×64 area, generating the selected areas image depicted in Figure 3.4 (selected areas). Finally, the selected areas are multiplied by the DCT diff Image. To better visualize the detected areas images that show the areas identified as having packet-loss artifacts, Figure 3.5 (detected areas) shows this result. For comparison, Figure 3.6 depicts the original impaired frame in question, what allow us to compare the areas containing packet-loss and the areas detected by the algorithm as having packet-loss. For instance, areas 1 and 2 in Figure 3.6 represent areas where the packet-losses can be found, while areas 3 and 4 in the same figure show areas with camera movement that generated strong edges.

The packet-loss measurement stage estimates the percentual strength of the packet-loss artifacts, detected in the previous stage, by extracting additional features from the blocks identified as packet-losses. A total of six types of features are extracted from the affected blocks: the sum of DC Energy (S_{DC}), the average of DC Energy (A_{DC}), the sum of the absolute value the first five AC coefficients (S_{AC}), the sum of horizontal AC coefficients (SH_{AC}), the sum of vertical AC coefficients (SV_{AC}) and difference of Borders (DB). To extract these features, the final image obtained from the detection stage is simultaneously split into blocks of 8×8, 16×16, and 32×32 pixels. The six types of features are processed with the 3 different block sizes, generating a total of 18 features.

Figure 3.7 illustrates how the DCT coefficient features are computed for a 8×8 block. The same structure (with adaptations) is used for 16x16 and 32x32 blocks. Notice that the block has

only one DC component, represented in green in Figure 3.7, which carries the block energy. The AC coefficients are used in the features S_{AC}, SH_{AC}, and SV_{AC}. The feature S_{AC} is calculated by summing the absolute values of the first five low-frequency elements, represented in yellow in Figure 3.7. The feature SH_{AC} is computed by summing the absolute values in the horizontal direction, represented in blue in Figure 3.7. The feature SV_{AC} is computed by summing the absolute values in the vertical direction, represented in red in Figure 3.7.

The feature DB differs from the other features because it requires the current frame and the detection image to calculate it. More specifically, when a block with packet-loss is identified, the difference between the intensity of pixels (or the coefficients) in the top and bottom of the borders is summed. Figure 3.8 illustrates these borders. The detection image shows the areas in which packet-loss artifacts were identified. Therefore, the differences between the top and bottom pixels of the borders are computed only for these areas.

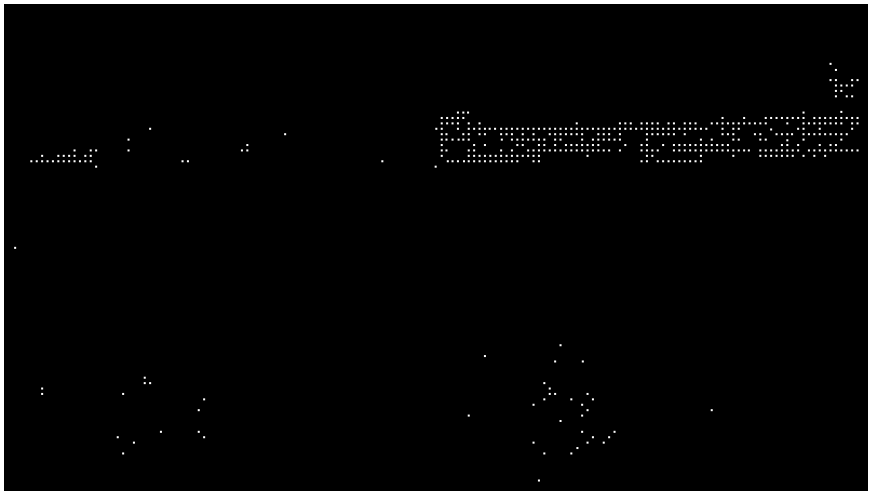


Figure 3.3: Picture displaying points in the frame selected as edges.

Figure 3.4: Picture showing the 64×64 areas selected as having packet-loss areas.

Figure 3.5: Detected Packet-Poss artifacts.

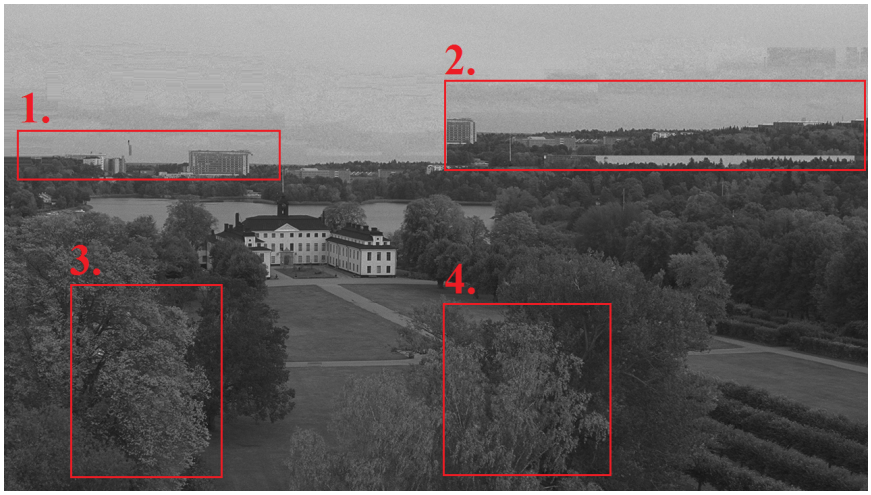


Figure 3.6: Database Varium Video 7 Frame 81, containing packet-loss artifacts (I=12 PLR=8.1%).

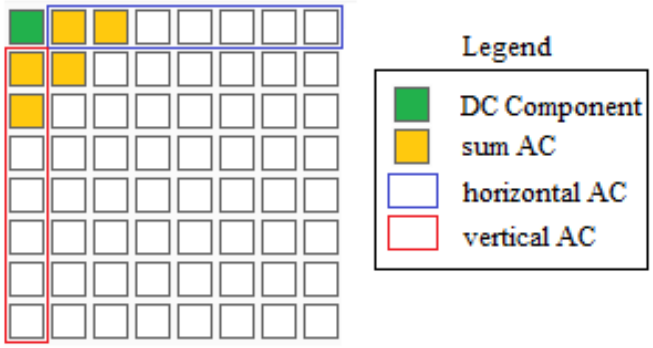


Figure 3.7: Block 8x8 - AC and DC features.

The feature DB is computed using the following equation:

$$DB = \sum_{i=1}^{NC} |B(i) - A(i)| + |C(i) - D(i)| \tag{3.7}$$

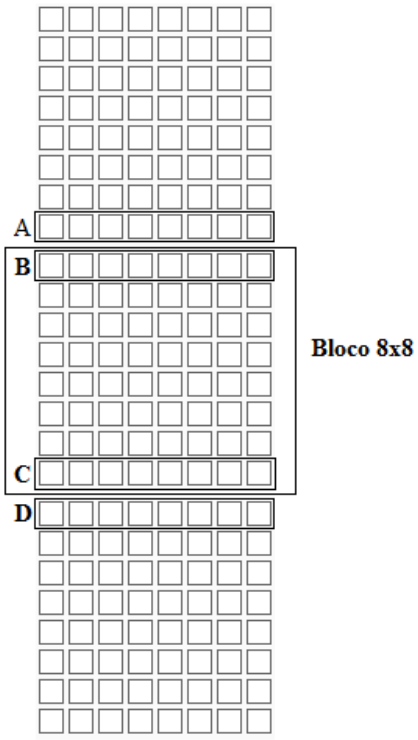


Figure 3.8: Difference Borders feature.

where B and C are 'up' borders, A and D are 'bottom' borders of the current block, and NC is the number of columns of the block (NC =8, 16, or 32 pixels).

The features S_{DC} and A_{DC} are computed in a similar way as DC Energy feature. The difference between them is basically that the S_{DC} feature is obtained by summing all DC Energy components while A_{DC} is obtained by taking the average of DC Energy of the selected blocks. The features S_{AC}, SH_{AC}, SV_{AC} and DB are obtained by counting the characteristics of all blocks affected by packet-loss. Each kind of block (8x8,16x16,32x32) creates a set features.

Figure 3.9 depicts a block diagram of the Feature Extration procedure. Notice that a total of eighteen features are obtained by the extraction procedure. However, not all eighteen features are statically relevant. Experimental results show that only three features were found to be statistically significant in the temporal analysis: $A_{DC,32}$, DB_{32}, $SV_{AC,32}$. A SVR algorithm combines these features in an optimal way, taking into consideration the subjective quality values (MOS) of the test sequences provided by the quality databases.

The block diagram of spatial analysis (Figure 3.2) is performed in a similar way, with the only difference being that the spatial input uses a sobel filter in the current frame. The computation of the DCT, detection stage, image with selected pixels, image with selected area and image with detection images follow the same logic of the temporal analysis. SVR tests performed found that three features are statistically significant: $SB\_S_{AC,16}$, SB_DB_{32} and $SB\_SV_{AC,8}$.

Figure 3.10 illustrates all packet-loss features, which include the spatial and temporal analysis.

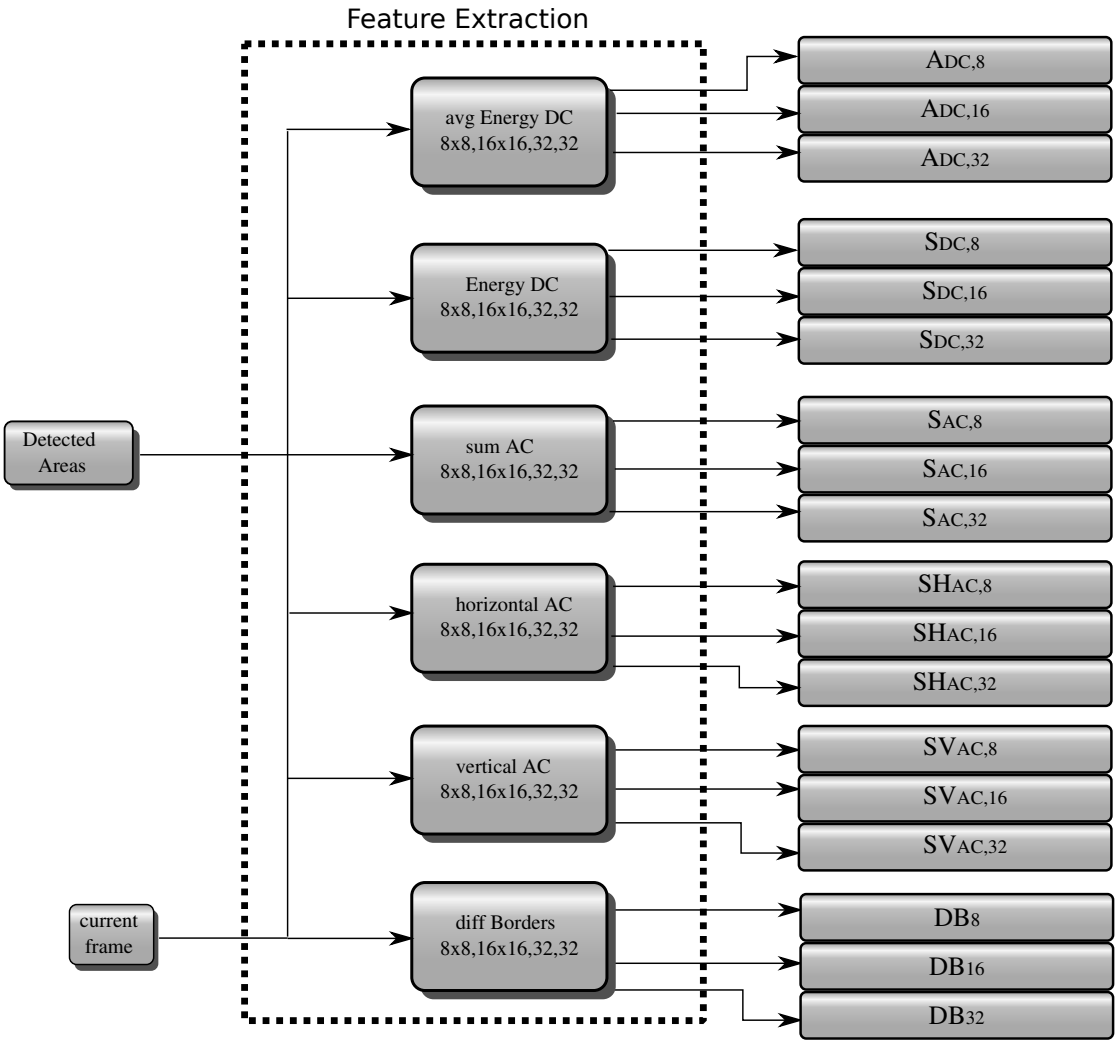


Figure 3.9: Measurement stage Feature Extraction.

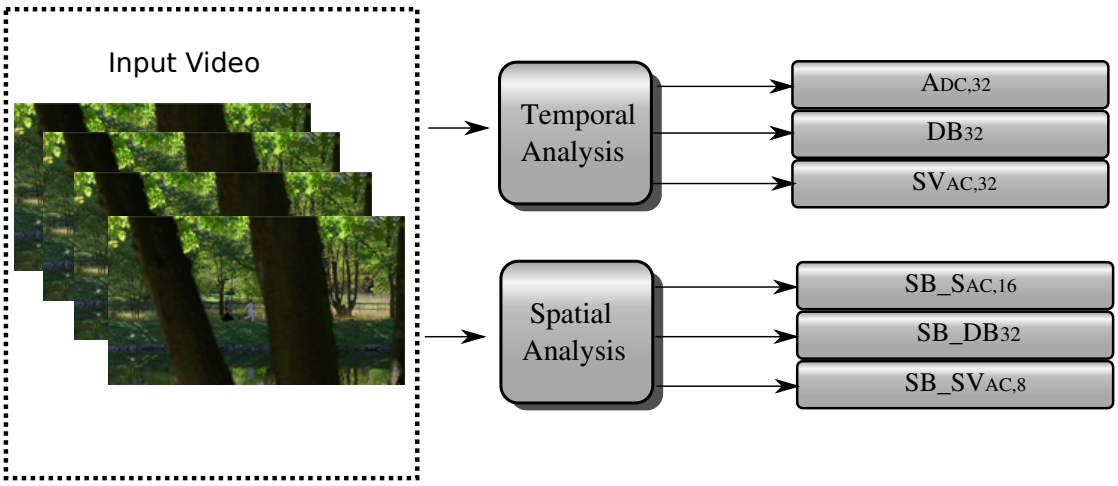


Figure 3.10: Packet-loss Summary.

Notice that three temporal and three spatial features are used. These six resulting features are going to be used to compose to hybrid metric that will be explained in the next section.

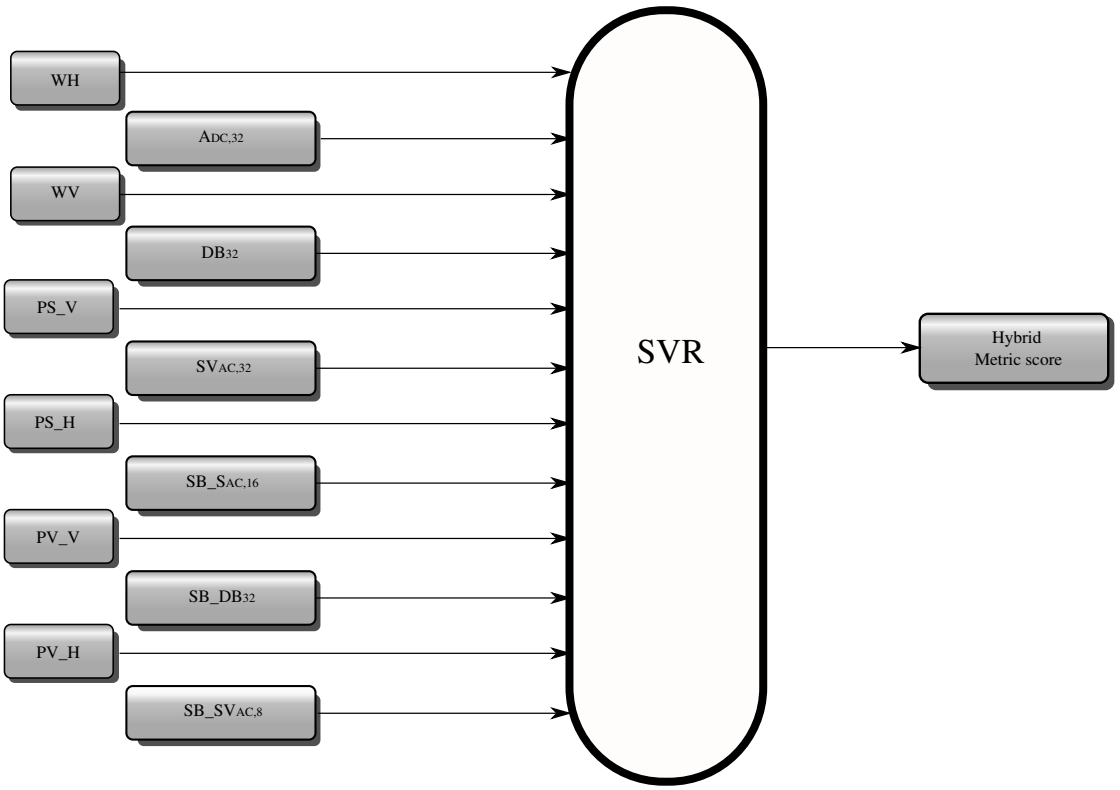


Figure 3.11: Hybrid Estimator.

3.3 HYBRID METRIC

The proposed hybrid metric combines several features that are taken from the artifact metrics presented earlier to estimate the overall quality of the video. These features are combined with an SVR technique taking into account three types of artefacts: packet-loss, bluriness and blockiness.

The six packet-loss features are obtained as explained in block diagram in Figure 3.10. The blockiness features are obtained by the equations 2.9 and 2.10. The bluriness features of each frame are based on equations 2.12, 2.13, 2.14 and 2.15 and the video features are obtained using the following equations:

$$PS\_V = \sum_{k=1}^{NF} ID_H(k) \tag{3.8}$$

$$PS\_H = \sum_{k=1}^{NF} ID_v(k) \tag{3.9}$$

$$PV\_V = \sum_{k=1}^{NF} MD_H(k) \tag{3.10}$$

$$PV\_H = \sum_{k=1}^{NF} MD_V(k) \tag{3.11}$$

Figure 3.11 shows a block diagram of hybrid process. In total, there are twelve features. Six features correspond to the packet-loss artifact ($A_{DC,32}$, DB_{32}, $SV_{AC,32}$, $SB\_S_{AC,16}$, SB_DB_{32}, $SB\_SV_{AC,8}$), two features correspond to the blockiness artifact (WH, WV) and four correspond to bluriness artifact (PS_V, PS_H, PV_V, PV_H). It is worth pointing out that we take into account only features metrics and not the overall score provided by these metrics.

4 RESULTS

In this chapter, we present the results of the proposed hybrid NR video quality metric. Tests are performed using five video quality databases: Varium, Roma, Live, CSIQ and IVPL (see Section 2.4). We compare the results of our proposed methodology with two NR packet-loss metrics and one FR quality metric.

4.1 TESTS OF INDIVIDUAL FEATURES

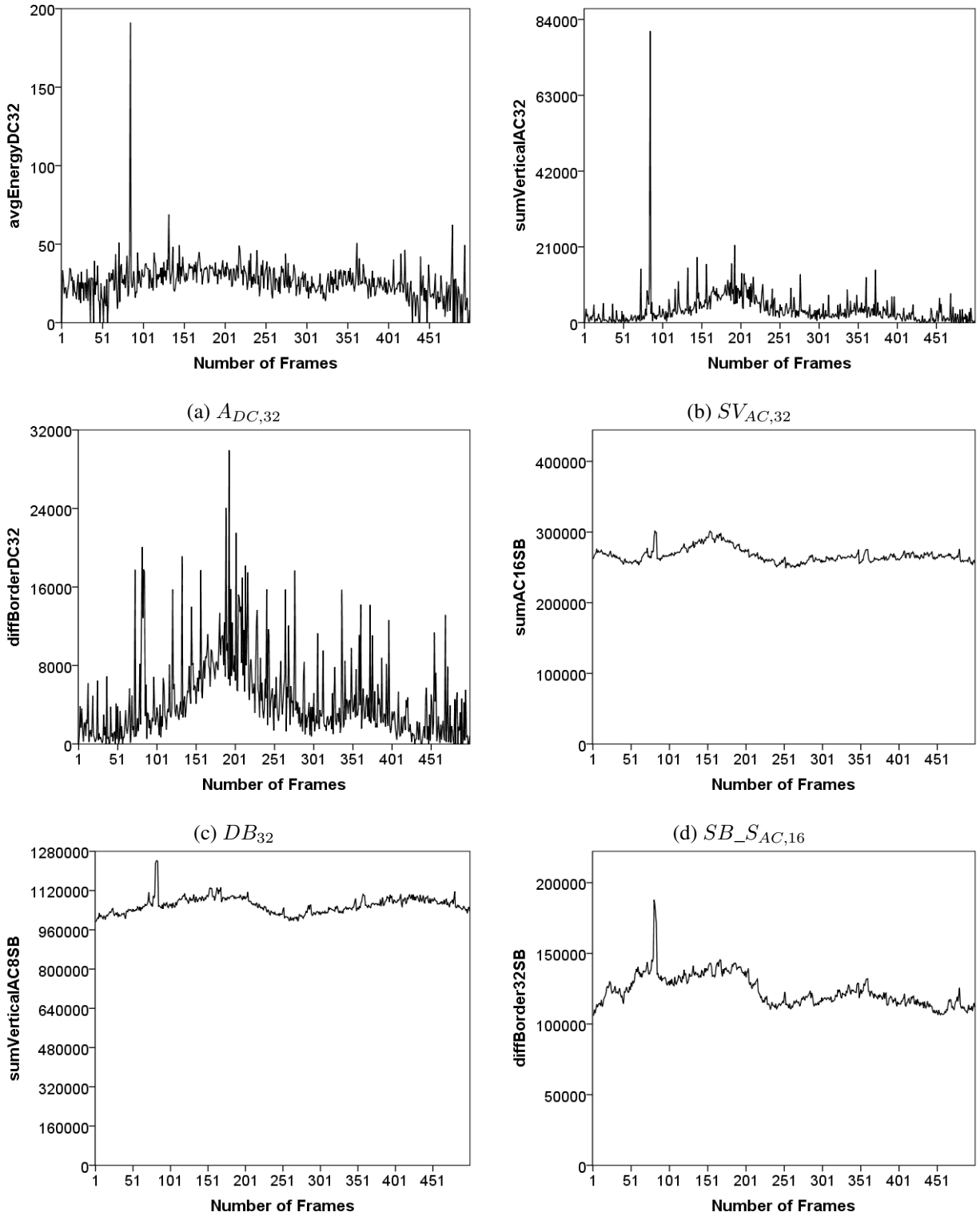
This section presents the results of each feature individually. First, the outcomes of the six packet-loss features are shown in a video with strong packet-loss artifact. As the packet-loss metric uses a new methodology, it is necessary to show if the features quantify well the strength of the artifacts. Second, the twelve features of the proposed hybrid metric are submitted under the five video quality databases to verify the behavior of them.

As explained in the previous chapter, twelve is the number of extracted features. Six are originated from packet-loss metric ($A_{DC,32}$, DB_{32}, $SV_{AC,32}$, $SB\_S_{AC,16}$, SB_DB_{32}, $SB\_SV_{AC,8}$), two from a blockiness metric (WH and WV) and four from a bluriness metric (PS_V, PS_H, PV_V, PV_H). Figures 4.1, 4.2 and 4.3 show the behavior of each feature individually. These graphs are computed for a video of Varium database containing packet-loss distortions at a packet-loss-rate (PLR) equals to 8.1% and a duration of 12 frames. Moreover, features are computed for each of the 500 frames of the video.

Figure 4.1 depicts the packet-loss feature response. The six features are tested in a video with strong packet-loss artifacts, located between frames 80 and 95. Notice that all features designed for packet-loss are able to identify these distortions. Figure 4.1 (a) shows the feature $A_{DC,32}$, while Figure 4.1 (b) shows the $SV_{AC,32}$. Both graphs have strong peaks near these frames. Figure 4.1 (d) shows $SB\_S_{AC,16}$, Figure 4.1 shows (e) $SB\_SV_{AC,8}$ and Figure 4.1(f) shows SB_DB_{32}. Again, notice that they all show a peak around the mentioned frames. Figure 4.1 (c) shows DB_{32}, which also shows a peak near the frames where the packet loss occurs. Nevertheless, the graph has other peaks which shows the existence of strong edge areas in the frame and it is not necessarily packet-loss artifacts.

Plots in Figure 4.2 depict the responses of blockiness features, while the plots in Figure 4.3 depict the responses of bluriness features. Observe that, none of the six features show a clear detection of the packet-loss artifacts. This result is expected since this video is affected *only* with packet-loss and it does not contain blockiness or bluriness artifacts.

Next, let's see the behaviour of the twelve features in the five datatabases. Tables from 4.1 to 4.6 show the correlation values in the databases: CSIQ, Live, IVPL, Roma, and Varium. Table 4.1



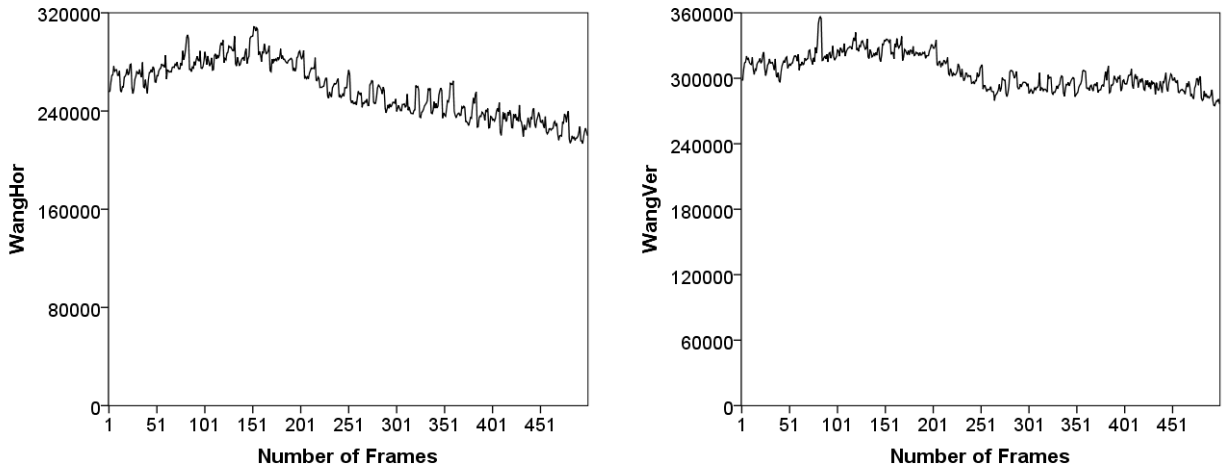
(a) $A_{DC,32}$ (b) $SV_{AC,32}$

(c) DB_{32} (d) $SB\_S_{AC,16}$

(e) $SB\_SV_{AC,8}$ (f) SB_DB_{32}

Figure 4.1: Responses of the Packet-Loss Features to a video with strong packet-loss artifacts, located between frames 80 and 95.

shows the Pearson (PCC) and Spearman (SCC) correlation coefficient values [47] for the CSIQ database. The highest value of correlation is -0.51 obtained for the $A_{DC,32}$ feature for the HEVC

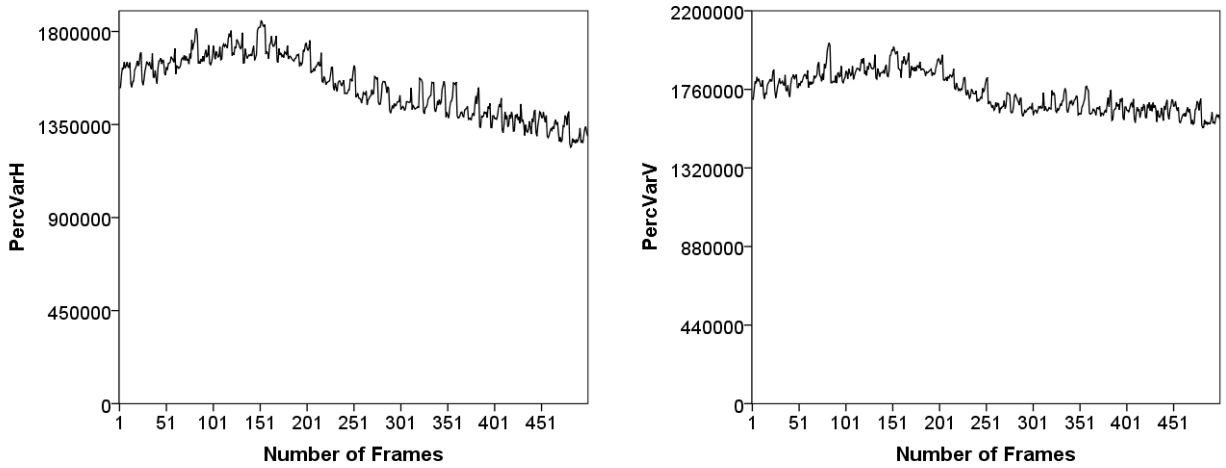


(a) WH (b) WV

Figure 4.2: Responses of Blockiness Features to a video with strong packet-loss artifacts, located between frames 80 and 95.

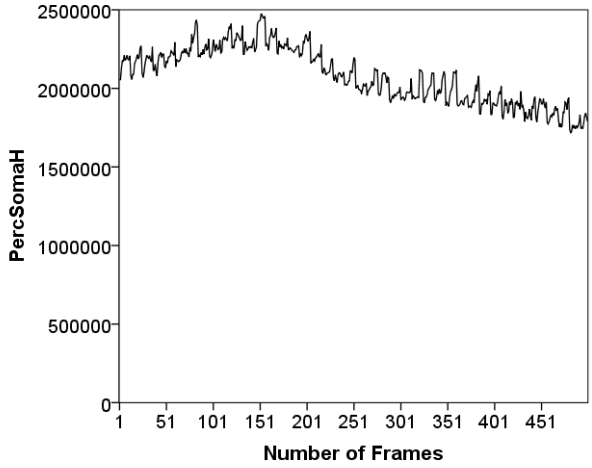
distortion. Table 4.2 indicates the PCC and SCC correlation values for the Live database. The maximum average correlation values are around 0.30 and the peak value 0.475 corresponds to the PV_H feature for the MPEG distortion. Table 4.3 shows the PCC and SCC values for the IPVL database. In this case, the maximum correlation values are a little higher, i.e the peak value is 0.587 corresponding to the PV_V feature for the H264 artifact. Table 4.4 shows the PCC and SCC values for the Roma and Varium Set 1, respectively. These databases have mainly packet-loss artifacts. In the Roma database, the maximum peak correlation values obtained is -0.338 for the WV feature. In the the Varium Set 1 the maximum is 0.560 for the $SB\_SV_{AC,8}$ feature. On average, the correlation values obtained for the Roma database are smaller than Varium Set 1. Table 4.5 shows the PCC and SCC values for the Varium Set 2. The correlation maximum peak achieved is 0.800 for the PS_H feature. Table 4.6 shows the PCC and SCC values for the Varium Set 3. The correlation values in this table are very small. Thus, each feature cannot be represented for these values in this database.

As a conclusion, individual features do not have a high correlation values with subjectives quality scores given by human observers. This, however, does not mean that their combination cannot provide a good prediction of the overall quality.

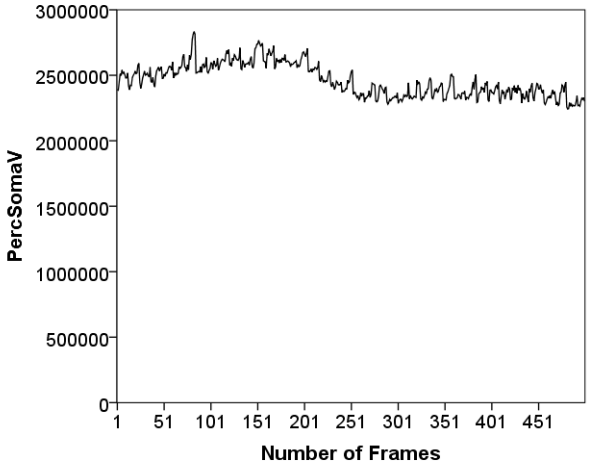


(a) PV_H

(b) PV_V



(c) PS_H



(d) PS_V

Figure 4.3: Responses of Bluriness Features to a video with strong packet-loss artifacts, located between frames 80 and 95.

Table 4.1: CSIQ - Pearson (PCC) and Spearman (SCC) correlation coefficients.

CSIQ												
Feature	H264		Pack		MJPEG		Wavelet		Noise		HEVC	
	PCC	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**
$A_{DC,32}$	0.120	0.010	-0.290	-0.300	0.106	0.130	-0.090	-0.040	-0.196	-0.200	**-0.510**	**-0.380**
DB_{32}	-0.238	-0.130	**-0.255**	**-0.310**	-0.128	-0.390	-0.163	-0.070	0.090	0.050	-0.139	-0.090
$SV_{AC,32}$	-0.019	-0.300	-0.069	-0.060	0.192	0.140	-0.080	-0.210	-0.191	-0.290	**-0.261**	**-0.110**
$SB\_S_{AC,16}$	-0.079	-0.180	0.246	0.250	-0.091	0.020	**-0.360**	**-0.350**	-0.120	0.080	0.180	0.280
SB_DB_{32}	**-0.283**	**-0.250**	0.058	0.010	0.128	0.110	-0.213	-0.210	0.256	0.120	-0.093	-0.080
$SB\_SV_{AC,8}$	-0.085	0.020	0.117	0.020	**0.255**	**0.190**	-0.151	-0.110	-0.092	-0.180	-0.115	-0.140
WH	0.105	0.190	0.215	-0.010	-0.349	-0.350	-0.486	-0.310	**0.506**	**0.460**	0.191	0.230
WV	-0.110	-0.080	0.148	0.170	0.328	0.440	-0.180	-0.210	**0.340**	**0.160**	-0.037	0.010
PV_H	-0.011	-0.070	0.120	0.070	-0.455	-0.470	-0.292	-0.230	**0.488**	**0.350**	0.201	0.370
PV_V	-0.029	-0.030	0.177	0.030	0.255	0.080	-0.021	0.100	**0.302**	**0.200**	0.125	0.170
PS_H	0.305	0.250	0.225	-0.050	-0.347	-0.380	-0.251	-0.220	**0.498**	**0.360**	0.234	0.320
PS_V	-0.085	-0.060	0.154	0.150	0.059	0.010	-0.016	-0.150	**0.187**	**0.180**	0.001	0.020

Table 4.2: Live - Pearson (PCC) and Spearman (SCC) correlation coefficients.

Live								
Feature	Wireless		IP		H264		MPEG2	
	PCC	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**
$A_{DC,32}$	0.079	0.220	0.269	0.200	-0.240	-0.140	**-0.416**	**-0.280**
DB_{32}	0.156	0.220	**-0.324**	**-0.350**	0.226	0.220	-0.190	-0.220
$SV_{AC,32}$	**0.289**	**0.300**	0.102	0.300	-0.149	-0.200	-0.171	-0.160
$SB\_S_{AC,16}$	0.241	0.360	0.102	0.150	-0.176	-0.140	**-0.258**	**-0.140**
SB_DB_{32}	0.232	0.260	0.212	0.100	-0.161	-0.020	**-0.297**	**-0.260**
$SB\_SV_{AC,8}$	0.184	0.140	-0.320	-0.350	-0.106	-0.160	**-0.404**	**-0.280**
WH	0.277	0.320	0.039	-0.050	-0.017	0.060	**-0.307**	**-0.240**
WV	0.293	0.300	0.011	0.000	-0.284	-0.160	**-0.404**	**-0.340**
PV_H	0.127	0.060	-0.304	-0.250	0.215	0.140	**0.475**	**0.440**
PV_V	0.216	0.260	**-0.376**	**-0.250**	-0.249	-0.340	-0.295	-0.260
PS_H	**0.287**	**0.320**	-0.282	-0.350	0.030	0.080	0.037	-0.100
PS_V	0.258	0.240	0.026	-0.150	-0.351	-0.280	**-0.364**	**-0.300**

Table 4.3: IVPL - Pearson (PCC) and Spearman (SCC) correlation coefficients.

IVPL						
Feature	Dirac		H264		MPEG2	
	PCC	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**
$A_{DC,32}$	0.353	0.200	**0.418**	**0.180**	-0.138	-0.150
DB_{32}	**0.248**	**0.300**	0.140	-0.040	-0.122	-0.100
$SV_{AC,32}$	**-0.273**	**-0.050**	-0.181	-0.240	-0.166	-0.150
$SB\_S_{AC,16}$	0.114	0.100	0.**420**	**0.480**	-0.392	-0.350
SB_DB_{32}	0.280	0.450	**0.499**	**0.540**	0.083	0.050
$SB\_SV_{AC,8}$	0.044	0.050	**0.546**	**0.560**	-0.164	-0.100
WH	0.227	0.100	0.328	0.480	**-0.391**	**-0.350**
WV	-0.084	-0.150	**0.541**	**0.480**	-0.007	0.200
PV_H	**0.458**	**0.450**	0.312	0.380	0.079	0.150
PV_V	0.085	0.150	**0.587**	**0.520**	0.027	0.100
PS_H	0.266	0.200	0.273	0.480	**-0.509**	**-0.250**
PS_V	-0.015	-0.050	**0.512**	**0.400**	0.106	0.200

4.2 HYBRID NR VIDEO QUALITY METRIC

This section explains how tests are performed and shows the results of the hybrid proposed metric.

To train and test the proposed metric, a k-fold cross validation setup is used, which consists of splitting the dataset up in k equally sized, non-overlapping sets. Then, we perform the test k times, in each time a different groups (fold) are used as the test set, and the remaining $k-1$ folds are used for training. This way, each data point has a chance of being validated against the other [46]. In our experiments, k is set to 10, thereby running 10 repetitions of the training. In terms of videos, it represents 90% for training and 10% for test. The k-fold cross validation setup is used to avoid the overfitting of the regression model. After predicting the automatic score, a

Table 4.4: Roma and Varium Set 1 - Pearson (PCC) and Spearman (SCC) correlation coefficients.

Roma and Varium Set 1				
	Roma		Set 1	
Feature	Pkt Loss		Pkt Loss	
	PCC	**SCC**	**PCC**	**SCC**
$A_{DC,32}$	0.049	0.110	0.559	0.501
DB_{32}	0.004	-0.020	0.471	0.477
$SV_{AC,32}$	0.164	0.109	0.491	0.441
$SB\_S_{AC,16}$	0.138	0.141	0.529	0.533
SB_DB_{32}	0.161	0.153	0.533	0.543
SB_SV_{AC}	-0.069	-0.079	0.560	0.541
WH	-0.015	0.088	0.515	0.538
WV	-0.338	-0.266	0.547	0.519
PV_H	-0.237	-0.185	0.440	0.443
PV_V	-0.283	-0.206	0.440	0.420
PS_H	-0.005	0.143	0.546	0.524
PS_V	-0.153	-0.127	0.537	0.490

Table 4.5: Varium Set 2 - Pearson (PCC) and Spearman (SCC) correlation coefficients.

Varium - Set 2						
Feature	Bloc		Blur		BlocBlur	
	PCC	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**
$A_{DC,32}$	-0.043	-0.050	**-0.332**	**-0.350**	-0.222	-0.200
DB_{32}	0.027	0.050	**-0.398**	**-0.400**	0.125	0.150
$SV_{AC,32}$	0.025	0.050	**0.113**	**0.050**	-0.111	-0.150
$SB\_S_{AC,16}$	**-0.400**	**-0.400**	0.008	0.000	-0.257	-0.300
SB_DB_{32}	-0.198	-0.200	**0.215**	**0.250**	-0.079	-0.100
$SB\_SV_{AC,8}$	**-0.284**	**-0.250**	-0.191	-0.150	0.087	-0.050
WH	**-0.600**	**-0.600**	-0.181	-0.150	-0.082	-0.150
WV	-0.196	-0.200	**-0.199**	**-0.200**	-0.131	-0.150
PV_H	-0.196	-0.200	-0.087	-0.150	**-0.208**	**-0.150**
PV_V	**-0.394**	**-0.400**	-0.200	-0.200	-0.272	-0.350
PS_H	**-0.800**	**-0.800**	-0.188	-0.150	0.026	0.000
PS_V	0.079	0.050	**-0.193**	**-0.150**	-0.029	-0.100

correlation is computed between the subjective data (MOS) and the predicted scores.

Table 4.9 shows the correlation values of the proposed NR Video Quality Hybrid Metric. For comparison purposes, we also show the results for the metrics Babu [20], Xia Rui [48], and SSIM [25]. Tables 4.7 and 4.8 show the correlation values per distortion type. There are nine types of distortions: H.264, PktLoss, MJPEG, Noise HEVC, Wireless, IP, MPEG2, Dirac, and the seven datasets. The CSIQ database has five distortions types: H.264, PktLoss, MJPEG, Noise, HEVC. The correlation values for CSIQ are around of 0.5. The Live, in turn, has four distortions: H264, Wireless, IP, MPEG2. The correlation values for the Live database range from 0.310 to 0.411. IVPL has three types of distortions: H.264, MPEG2 and Dirac. The correlation values for the IVPL database range from 0.388 to 0.526.

Table 4.6: Varium Set 3 - Pearson (PCC) and Spearman (SCC) correlation coefficients.

Varium - Set 3												
Feature	Pack		Bloc		Blur		PackBloc		PackBlur		PackBlocBlur	
	PCC	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**
$A_{DC,32}$	-	-	-	-	-	-	**-0.452**	**-0.400**	-0.046	-0.150	0.228	0.228
DB_{32}	-	-	-	-	-	-	-0.034	-0.100	-0.217	-0.100	**0.442**	**0.483**
$SV_{AC,32}$	-	-	-	-	-	-	-0.098	-0.200	0.065	0.000	**0.509**	**0.540**
$SB\_S_{AC,16}$	-	-	-	-	-	-	**-0.461**	**-0.300**	-0.077	0.050	0.205	0.311
SB_DB_{32}	-	-	-	-	-	-	-0.355	-0.250	-0.077	0.250	**0.370**	**0.523**
$SB\_SV_{AC,8}$	-	-	-	-	-	-	-0.233	-0.200	0.317	0.450	**0.577**	**0.621**
WH	-	-	-	-	-	-	-0.172	-0.050	-0.214	-0.150	**0.377**	**0.406**
WV	-	-	-	-	-	-	-0.029	0.100	-0.278	-0.250	**0.313**	**0.377**
PV_H	-	-	-	-	-	-	-0.262	-0.300	**-0.268**	**-0.300**	0.198	0.140
PV_V	-	-	-	-	-	-	**-0.411**	**-0.350**	-0.116	-0.100	0.316	0.326
PS_H	-	-	-	-	-	-	-0.134	-0.100	0.031	-0.050	**0.269**	**0.203**
PS_V	-	-	-	-	-	-	-0.588	-0.550	-0.255	-0.150	**0.326**	**0.284**

Table 4.7: Pearson (PCC) and Spearman (SCC) correlation coefficients per distorton for the proposed hybrid metric - part 1.

Distortion	CSIQ		Live		IVPL		Roma		Set 1		Set 2 Bloc		Set 2 Blur	
	PCC	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**
H.264	0.542	0.569	0.385	0.375	0.440	0.388	-	-	-	-	0.802	0.794	0.677	0.617
PktLoss	0.479	0.541	-	-	-	-	0.709	0.713	0.790	0.770	-	-	-	-
MJPEG	0.529	0.516	-	-	-	-	-	-	-	-	-	-	-	-
Noise	0.521	0.554	-	-	-	-	-	-	-	-	-	-	-	-
HEVC	0.542	0.547	-	-	-	-	-	-	-	-	-	-	-	-
Wireless	-	-	0.409	0.352	-	-	-	-	-	-	-	-	-	-
IP	-	-	0.411	0.356	-	-	-	-	-	-	-	-	-	-
MPEG2	-	-	0.334	0.310	0.491	0.401	-	-	-	-	-	-	-	-
Dirac	-	-	-	-	0.526	0.438	-	-	-	-	-	-	-	-

Notice that the proposed metric has better results than the packet-loss metrics. The correlation values for Roma database are around 0.700, what represents a good performance. The best correlation values of Tables 4.7 and 4.8 are for the Varium Sets 1, 2 and 3. Table 4.7 Set 1, only with the packet-loss artifacts, shows correlations values of 0.790. Table 4.8 Set 2, with blockiness and bluriness, shows maximum correlation of 0.802 and Set 3 ,with blockiness, bluriness and packet-loss, shows maximum correlation of 0.842.

Table 4.9 depicts the comparison between the proposed metric and two other no-reference packet loss metrics and full reference metric. From the results, the proposed metric shows a better correlation in almost all databases. The SSIM metric [49] has higher correlation values in two databases: CSIQ and Live. Even though SSIM is a full reference metric, the proposed metric correlation values for these two databases are not very different for SSIM values. For other databases the proposed metric has a higher performance. The NR packet-loss metrics, Babu and Xia Rui, have much lower correlation values than the proposed metric. Except in the case of the Roma database, in which Xia Rui and proposed metrics show similar correlation values. But the proposed metrics have a better prediction accuracy.

Table 4.8: Pearson (PCC) and Spearman (SCC) correlation coefficients per distorton for the proposed hybrid metric - part 2.

Distortion	Set 2 BlocBlur		Set 3 Bloc		Set 3 Blur		Set 3 PLR		Set 3 PLRBloc		Set 3 PLRBlur		Set 3 PLRBlocBlur	
	PCC	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**	**PCC**	**SCC**
H.264	0.790	0.748	0.804	0.807	0.835	0.842	-	-	0.821	0.813	0.831	0.829	0.820	0.828
PktLoss	-	-	-	-	-	-	0.764	0.783	-	-	-	-	-	-
MJPEG	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Noise	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HEVC	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Wireless	-	-	-	-	-	-	-	-	-	-	-	-	-	-
IP	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MPEG2	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Dirac	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 4.9: Comparison of correlation coefficients per reference metrics.

Database	CSIQ		LIVE		ROMA		IVPL		Set 1		Set 2		Set3	
	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC
Proposed	**0.525**	**0.530**	**0.452**	**0.452**	**0.691**	**0.658**	**0.452**	**0.452**	**0.758**	**0.750**	**0.736**	**0.726**	**0.846**	**0.866**
Babu	0.081	0.085	-0.035	0.015	-0.161	-0.274	-0.012	0.042	0.352	0.318	-0.142	-0.108	0.072	-0.068
Xia Rui	0.126	0.088	0.127	0.137	0.678	0.587	0.222	0.133	0.389	0.369	0.313	0.327	0.475	0.493
SSIM	0.660	0.635	0.507	0.471	0.252	0.201	0.069	0.106	0.622	0.542	0.430	0.395	0.689	0.642

5 CONCLUSIONS

In this dissertation, we proposed a hybrid no-reference video quality metric, which considers twelve features obtained from packet-loss, blockiness and bluriness metrics. The packet loss features are achieved in two steps (detection and measure), generating a total of six features. Three features are temporal (A_{DC32}, DB_{32}, $SV_{AC,32}$) and the other three are spatial features ($SB\_S_{AC,16}$,SB_DB_{32}, $SB\_SV_{AC,8}$). The two blockiness features (WH and WV) come from Wang's algorithm and the four blurinesss features (PV_V, PV_H, PS_V, PS_H) from Crété-Roffet's algorithm.

When each feature was evaluated individually, results were not satisfactory with almost all correlation values are below 0.5. However, when features were combined to compose a hybrid NR video quality metric, outcomes improved considerably. The hybrid metric shows correlation values higher than 0.7 in almost all databases where packet-loss, blockiness and bluriness are present.

Tests were performed using five video quality databases, which contained nine types of distortions: H.264, PktLoss, MJPEG, Noise HEVC, Wireless, IP, MPEG2, Dirac. If you consider only the databases that contain packet-loss, bluriness and blockiness (Roma, Varium Set 1, Varium Set 2 and Varium Set 3), the correlation results show a satisfatory performance (approximately 0.7 or higher). On the other hand, for the other databases, the correlation values were lower than 0.57. This can be explained by the fact that these databases contain certain types of distortions that are not considered by the proposed metric.

A comparison with other reference metrics was performed to test the performance of the proposed metric. The SSIM metric did not behave as expected, showing a high performance for only two databases: CSIQ and Live. The other NR packet-loss metrics (Xia Rui and Babu) also showed very low correlation values. In summary, the proposed hybrid metric shows a higher performance than the tested FR and NR metrics, for the databases Roma, IVPL, and Varium Sets 1, 2, 3.

5.1 FUTURE WORKS

As future work, other types of artifact features can be included in the design of the hybrid metric. Examples include noise, ringing, color degradations, etc. We believe that having a more diverse set of features can improve the correlation. Furthermore, the performance of the packet-loss detection algorithm needs to be improved to reduce the rate of false positives. As the DCT worked quite well for detecting packet-loss, a possibility would be to extract blockiness and bluriness features from the same DCT frame already computed. Thus, packet-loss, blockiness, and bluriness features would be extracted from just one DCT frame, what would reduce the computational cost of the proposed metric.

In terms of time performance, future works could include the implementation of SVR technique in a faster programming language to make it possible to run applications in real time. Also, DCT was applied in blocks 8x8. It would be interesting to test other block sizes (16x16, 32x32) to verify if results would improve. Finally, in this work features are extracted from all frames of a video. An interesting work would be to test the performance of the proposed method when we reduce the number of frames and the spatial resolution.

REFERÊNCIAS BIBLIOGRÁFICAS

1 PAPER, C. W. Cisco visual networking index: Global mobile data traffic forecast update, 2013-2018. 2014.

2 WINKLER, S.; MOHANDAS, P. The evolution of video quality measurement: from psnr to hybrid metrics. *Broadcasting, IEEE Transactions on*, IEEE, v. 54, n. 3, p. 660–668, 2008.

3 ITU. *BT.500-13, Methodology for the subjective assessment of the quality of television pictures*. [S.l.]: ITU, 2012.

4 MORAIS A.F. SILVA, M. Q. F. D. A correlation-based no-reference packet-loss metric. *XXXIV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*, set 2016.

5 WANG, Z.; LU, L.; BOVIK, A. C. Video quality assessment based on structural distortion measurement. *Signal processing: Image communication*, Elsevier, v. 19, n. 2, p. 121–132, 2004.

6 PINSON, M. H.; WOLF, S. A new standardized method for objectively measuring video quality. *Broadcasting, IEEE Transactions on*, IEEE, v. 50, n. 3, p. 312–322, 2004.

7 GROUP, V. Q. E. *Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment - Phase II*. [S.l.], 2003.

8 FARIAS, M. C.; MITRA, S. K. No-reference video quality metric based on artifact measurements. In: IEEE. *Image Processing, 2005. ICIP 2005. IEEE International Conference on*. [S.l.], 2005. v. 3, p. III–141.

9 LUBIN, J. *Final report [microform] : Sarnoff JND Vision Model for flat-panel design : contract no. NAS2-14257 : period of performance, Jan. 1996-Jan. 1998 / prepared by Visual Information Systems Research Group, Information Sciences Laboratory, Sarnoff Corporation*. [S.l.]: National Aeronautics and Space Administration ; National Technical Information Service, 1998.

10 GUNAWAN, I. P.; GHANBARI, M. Image quality assessment based on harmonics gain/loss information. In: *IEEE International Conference on Image Processing 2005*. [S.l.: s.n.], 2005. v. 1, p. I–429–32. ISSN 1522-4880.

11 KANUMURI, S.; SUBRAMANIAN, S. G.; COSMAN, P. C.; REIBMAN, A. R. Predicting h. 264 packet loss visibility using a generalized linear model. In: IEEE. *Image Processing, 2006 IEEE International Conference on*. [S.l.], 2006. p. 2245–2248.

12 MITTAL, A.; SOUNDARARAJAN, R.; BOVIK, A. C. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, IEEE, v. 20, n. 3, p. 209–212, 2013.

13 MOORTHY, A. K.; BOVIK, A. C. Blind image quality assessment: From scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, v. 20, n. 12, p. 3350–3364, Dec 2011.

14 ITU. *Recommendation E.800, Definitions of terms Related to Quality of Service*. [S.l.]: ITU, 2008.

15 MAIA HANI CAMILLE YEHIA, L. d. E. O. B. A concise review of the quality of experience assessment for video streaming. *Elsevier*, Computer Communications, 2015.

16 ITU. *P.910, Subjective video quality assessment methods for multimedia applications*. [S.l.]: ITU, 2008.

17 S. BARAKOVIC J., B. H. B. Qoe dimensions and qoe measurement of ngn services. TELFOR, 2010.

18 LIN, W.; KUO, C.-C. J. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, Elsevier, v. 22, n. 4, 2011.

19 GARCIA, M.; SCHLEICHER, R.; RAAKE, A. Impairment-factor-based audiovisual quality model for IPTV: influence of video resolution, degradation type, and content type. *EURASIP J. Image and Video Processing*, v. 2011, 2011. Disponível em: <http://dx.doi.org/10.1155/2011/629284>.

20 BABU, R. V.; BOPARDIKAR, A. S.; PERKIS, A.; HILLESTAD, O. I. No-reference metrics for video streaming applications. In: *International Workshop on Packet Video*. [S.l.: s.n.], 2004.

21 BABU, R. V.; PERKIS, A.; HILLESTAD, O. I. Evaluation and monitoring of video quality for uma enabled video streaming systems. *Multimedia Tools and Applications*, Springer, v. 37, n. 2, p. 211–231, 2008.

22 GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006. ISBN 013168728X.

23 HASKELL BARRY G., P. A. N. A. N. *Digital Video: An introduction to MPEG-2*. [S.l.]: Springer, 2002. ISBN 978-0-306-46982-4.

24 WANG, Z.; BOVIK, A. C. Mean squared error: love it or leave it? a new look at signal fidelity measures. *Signal Processing Magazine, IEEE*, IEEE, v. 26, n. 1, p. 98–117, 2009.

25 WANG, Z.; BOVIK, A. C.; SHEIKH, H. R.; SIMONCELLI, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, v. 13, n. 4, p. 600–612, April 2004. ISSN 1057-7149.

26 WANG, Z.; SHEIKH, H. R.; BOVIK, A. C. No-reference perceptual quality assessment of jpeg compressed images. In: *Proceedings. International Conference on Image Processing*. [S.l.: s.n.], 2002. v. 1, p. I–477–I–480 vol.1. ISSN 1522-4880.

27 CRETE-ROFFET, F.; DOLMIERE, T.; LADRET, P.; NICOLAS, M. The Blur Effect: Perception and Estimation with a New No-Reference Perceptual Blur Metric. In: *SPIE Electronic Imaging Symposium Conf Human Vision and Electronic Imaging*. San Jose, USA: [s.n.], 2007. Disponível em: <https://hal.archives-ouvertes.fr/hal-00232709/document>.

28 CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273–297, 1995. ISSN 1573-0565. Disponível em: <http://dx.doi.org/10.1007/BF00994018>.

29 KARATZOGLOU, A.; MEYER, D.; HORNIK, K. Support vector machines in r. *Journal of Statistical Software*, v. 15, n. 1, p. 1–28, 2006. ISSN 1548-7660. Disponível em: <https://www.jstatsoft.org/index.php/jss/article/view/v015i09>.

30 THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern Recognition, Fourth Edition*. 4th. ed. [S.l.]: Academic Press, 2008. ISBN 1597492728, 9781597492720.

31 VAPNIK, V. N. *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8.

32 SILVA, A. F.; FARIAS, M. C.; REDI, J. A. Annoyance models for videos with spatio-temporal artifacts. In: IN: 2016 EIGHTH INTERNATIONAL CONFERENCE ON QUALITY OF MULTIMEDIA EXPERIENCE (QOMEX), LISBON. [S.l.], 2016.

33 FARIAS, M. C.; HEYNDERICKX, I.; ESPINOZA, B. M.; REDI, J. Visual artifacts interference understanding and modeling (varium). In: *Seventh international workshop on video processing and quality metrics for consumer electronics*. [S.l.: s.n.], 2013. v. 1.

34 REDI, J.; HEYNDERICKX, I.; MACCHIAVELLO, B.; FARIAS, M. On the impact of packet-loss impairments on visual attention mechanisms. In: IEEE. *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*. [S.l.], 2013. p. 1107–1110.

35 PAUDYAL, P.; BATTISTI, F.; CARLI, M. *Study of the effects of video content on quality of experience*. 2015. 93960W-93960W-9 p. Disponível em: <http://dx.doi.org/10.1117/12.2083223>.

36 PAUDYAL, P.; BATTISTI, F.; CARLI, M. A study on the effects of quality of service parameters on perceived video quality. In: *2014 5th European Workshop on Visual Information Processing (EUVIP)*. [S.l.: s.n.], 2014. p. 1–6.

37 BATTISTI, F.; CARLI, M.; PAUDYAL, P. Qos to qoe mapping model for wired/wireless video communication. In: *2014 Euro Med Telco Conference (EMTC)*. [S.l.: s.n.], 2014. p. 1–6.

38 SESHADRINATHAN, K.; SOUNDARARAJAN, R.; BOVIK, A. C.; CORMACK, L. K. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, v. 19, n. 6, p. 1427–1441, June 2010. ISSN 1057-7149.

39 SESHADRINATHAN, K.; SOUNDARARAJAN, R.; BOVIK, A. C.; CORMACK, L. K. *A subjective study to evaluate video quality assessment algorithms*. 2010. 75270H-75270H-10 p. Disponível em: <http://dx.doi.org/10.1117/12.845382>.

40 VU, P. V.; CHANDLER, D. M. Vis3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electronic Imaging*, v. 23, n. 1, p. 013016, 2014. Disponível em: <http://dx.doi.org/10.1117/1.JEI.23.1.013016>.

41 IVP. *IVP Subjective Quality Video Database*. 2011. Disponível em: <http://ivp.ee.cuhk.edu.hk/research/database/subjective/>.

42 T., V. Detection of blocking artifacts in compressed video. *Electronics Letters*, IET, v. 36, n. 13, p. 1106–1108, 2000.

43 KUSSABA, H. T. M.; FARIAS, M. C. Blind estimation of blocking artifacts in digital videos. *Latin Display 2010*, 2010.

44 GASTALDO, P.; ZUNINO, R.; REDI, J. Supporting visual quality assessment with machine learning. *EURASIP Journal on Image and Video Processing*, Springer, v. 2013, n. 1, p. 1–15, 2013.

45 BHATTACHARYYA, K.; JAMADAGNI, H. S. Dct coefficient-based error detection technique for compressed video stream. In: *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*. [S.l.: s.n.], 2000. v. 3, p. 1483–1486 vol.3.

46 REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. In: *Encyclopedia of database systems*. [S.l.]: Springer, 2009. p. 532–538.

47 JOAQUIM, P.; MARQUES, S. Applied statistics using spss, statistica, matlab and r. *Springer Company USA*, p. 205–211, 2007.

48 RUI, H.-x.; LI, C.-r.; QIU, S.-k. Evaluation of packet loss impairment on streaming video. *Journal of Zhejiang University SCIENCE A*, Springer, v. 7, n. 1, p. 131–136, 2006.

49 WANG, Z.; SIMONCELLI, E. P.; BOVIK, A. C. Multiscale structural similarity for image quality assessment. In: IEEE. *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*. [S.l.], 2003. v. 2, p. 1398–1402.

50 BOVIK, A. C. *The Essential Guide to Video Processing*. 2nd. ed. [S.l.]: Academic Press, 2009. ISBN 0123744563, 9780123744562.

6 APPENDIX

6.1 DISCRETE COSINE TRANSFORM (DCT)

Discrete Cosine Transform is important to several applications that uses lossy compression used in image and video compression systems such as MPEG-1,MPEG-2 and MPEG-4 codecs. Essentially, the idea of this process is to transform the image in the pixel domain to the frequency domain according to the following equation [50].

$$F(u,v) = \frac{1}{4} C_u C_v \sum_{i=0}^{7} \sum_{j=0}^{7} f(i,j) cos(\frac{(2i+1)u\pi}{16}) cos(\frac{(2j+1)v\pi}{16}) \tag{6.1}$$

where i and j are the horizontal and vertical the block indices, and u and v are the horizontal and vertical spatial frequencies indices. The constants C_u and C_v have the following values:

$C_u = \frac{1}{\sqrt{2}}$ for u=0, $C_u = 1$, otherwise

$C_v = \frac{1}{\sqrt{2}}$ for v=0, $C_u = 1$, otherwise

In order to have a better interpretation of how the DCT coefficients are distributed in transformed frequency block, Figure 6.1 (a) shows the frequency distribution and Figure 6.1 (b), the block feature of DCT coefficients. There is only one DC coefficient per block which gives the block energy, all other coefficients are AC coefficients, i.e. they represent the energy of a particular spatial frequency in the signal. If AC coefficients are close to the DC coefficient, they correspond to low frequency. So, in the image Figure 6.1 (a) is possible to see low, medium and high frequencies. The Figure 6.1 (b) represents the DC, horizontal, vertical and diagonal components.

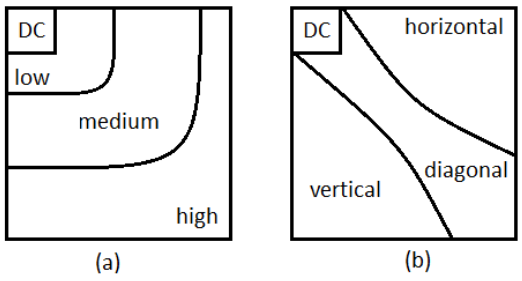


Figure 6.1: DCT - (a) Frequency distribution and (b) block features of DCT coefficients [45]

6.2 SOBEL

Sobel filters are simple and popular filters in image processing which are used to extract the edges of a image. More specifically, they are able to detect changes in intensity by taking gradient first deverivatives of the lines and columns of an image. The gradient is given by the following equation :

$$\nabla f \equiv grad(f) \equiv \begin{bmatrix} g_x \\ g_y \end{bmatrix} \equiv \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \tag{6.2}$$

where, f(x,y) is an image and (x,y) its coordinates. g_x and g_y are the gradient in the horizontal and vertical directions respectively. According to the above equation , the image is independently filtered with Gx and Gy.

$$gx = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{6.3}$$

$$gy = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{6.4}$$

The results are combined to obtain the final result. The magnitude, M, and the direction, θ, of the gradient are obtained using the following equations:

$$G = \sqrt{g_x^2 + gy^2} \tag{6.5}$$

$$\theta = \arctan[\frac{gy}{gx}] \tag{6.6}$$