DISSERTAÇÃO DE MESTRADO

ON THE PERFORMANCE OF VIDEO QUALITY ASSESSMENT METHODS FOR DIFFERENT SPATIAL AND TEMPORAL RESOLUTIONS

Welington Yorihiko Lima Akamine

Brasília, Fevereiro de 2017

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA
Faculdade de Tecnologia

DISSERTAÇÃO DE MESTRADO

ON THE PERFORMANCE OF VIDEO QUALITY ASSESSMENT METHODS FOR DIFFERENT SPATIAL AND TEMPORAL RESOLUTIONS

Welington Yorihiko Lima Akamine

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. Mylène Christine Queiroz de Farias, Ph.D, _____
ENE/UnB
Orientadora

Prof. Alexandre Ricardo Soares Romariz, Ph.D, _____
ENE/UnB
Examinador interno

Prof. Carlos Alexandre Barros de Mello, Ph.D, _____
CIn/UFPE
Examinador externo

FICHA CATALOGRÁFICA

AKAMINE, WELINGTON YORIHIKO LIMA
ON THE PERFORMANCE OF VIDEO QUALITY ASSESSMENT METHODS FOR DIFFERENT SPATIAL AND TEMPORAL RESOLUTIONS [Distrito Federal] 2017.
xvi, 78 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. Video Quality Metrics	2. Spatial Resolution
3. Temporal Resolution	4. Runtime Performance
I. ENE/FT/UnB	II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

AKAMINE, W. Y. L. (2017). *ON THE PERFORMANCE OF VIDEO QUALITY ASSESSMENT METHODS FOR DIFFERENT SPATIAL AND TEMPORAL RESOLUTIONS* . Dissertação de Mestrado , Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 78 p.

CESSÃO DE DIREITOS

Welington Yorihiko Lima Akamine
Depto. de Engenharia Elétrica (ENE) - FT
Universidade de Brasília (UnB)
Campus Darcy Ribeiro
CEP 70919-970 - Brasília - DF - Brasil

Acknowledgments

I would first like to acknowledge Mylène Farias, my tutor, for her patience, encouragement, and immense knowledge. Knowing that I could always was counted on her help and guidance, was what gave me the strength necessary to confront all adversity during this research. I could not be more grateful to have had her as my tutor.

I am grateful to my parents, Mercedes and Ruy, for their unconditionally love, help and support. And my sisters, Aline, Larissa and Alicia, for always being there for me as friends and supporting me during these years.

Also, I will be eternally thankful for the assistance given by Pedro Garcia. His valuable and constructive suggestions were essential during the development of this research work.

Finally, I thanks all my colleagues and friends in the Digital Signal Processing Group (GPDS). Gustavo and Johnathan with whom I shared a room at UnB. Dario, whom I met in my first day as a graduate student and I am thankful for his friendship during this journey. Helard and Alexandre, for helping me understand some of the algorithms used in this work.

Welington Yorihiko Lima Akamine

RESUMO

O consumo de vídeos digitais cresce a cada ano. Vários países já utilizam tv digital e o tráfego de dados de vídeos na internet equivale a mais de 60% de todo o tráfego de dados na internet. Esse aumento no consumo de vídeos digitais exige métodos computacionais viáveis para o cálculo da qualidade do vídeo.

Métodos objetivos de qualidade de vídeo são algoritmos que calculam a qualidade do vídeo. As mais recentes métricas de qualidade de vídeo, apesar de adequadas possuem um tempo de execução alto. Em geral, os algoritmos utilizados são complexos e extraem características espaciais e temporais ds vídeos.

Neste trabalho, realizamos uma análise dos efeitos da redução da resolução espacial no desempenho dos métodos de avaliação da qualidade do vídeo. Com base nesta análise, nós propomos um *framework*, para a avaliação da qualidade de vídeo que melhora o tempo de execução das métricas objetivas de qualidade de vídeo sem reduzir o desempenho na predição da qualidade do vídeo. O *framework* consiste em quatro etapas. A etapa de classificação identifica os vídeos mais sensíveis a redução da resolução espacial. A etapa de redução reduz a resolução espacial do vídeo de acordo com a distorção. A etapa da predição da qualidade é realizada por meio de uma métrica objetiva de qualidade. Finalmente, temos a etapa de ajuste dos índices de qualidade preditos.

Dois classificadores de vídeo são propostos para a primeira etapa do *framework*. O primeiro é um classificador com referência baseado na atividade espacial do vídeo. O segundo é um classificador sem referência baseado na entropia espacial e espectral, utilizando *Support Vector Machine* para classificar os vídeos. Os classificadores de vídeo têm um objetivo de selecionar o melhor fator de redução da resolução espacial do vídeo. Com isso, melhoramos o tempo de execução de todas as métricas de qualidade de vídeo testadas.

Também analisamos os efeitos da redução da resolução temporal no desempenho das métricas de qualidade de vídeo. A análise mostra que as métricas de qualidade de vídeo que se baseiam em características temporais são mais sensíveis à redução temporal. E, vídeos distorcidos por distorções temporais, como perda de pacote, são mais sensíveis à redução temporal.

ABSTRACT

The consume of digital videos is increasing every year. Many countries already use digital tv, and the traffic of internet video services are more than 60% of the total internet traffic. The growth of digital video consume demands a viable method to measure the video quality.

Objective video quality assessment methods are algorithms that estimates video quality. Recent video quality assessment methods are well correlated with subjective quality scores. However, most of these methods are very complex and takes long periods to compute.

A simple method to improve the running time of video quality assessment methods is reducing the video spatial resolution. We analyze the effects of spatial resolution reduction on the performance of video quality methods. The spatial resolution reduction decreases the accuracy performance of quality assessment methods for videos distorted by white noise, packet loss, and compressed with MJPEG. We propose a four stage framework, which has the goal of improving the runtime performance of video quality assessment methods without decreasing their quality estimation performance. The first stage identifies videos more sensitive to spatial resolution reduction. The second stage reduces the video spatial resolution according to the video distortion. The third stage estimates the video quality using an objective video quality method. Finally, the final stage adjusts the estimate quality score according to the video spatial resolution.

We design two video classifiers for the first stage of the framework. The first classifier is a full reference classifier based on a video spatial activity measure. The second is a no-reference classifier based on the spatial and spectral entropy, which uses a Support Vector Machine (SVM) algorithm. We use the video classifiers to choose the most appropriate video spatial resolution before computing the quality using a video quality assessment method. We tested the framework using six different video quality assessment methods. Results show that the proposed framework improves the average runtime performance of all video quality assessment methods tested.

Also, we analyze the effects of a temporal resolution reduction on the performance of video quality assessment methods. The analysis shows that video quality assessment methods based on temporal features are more sensitive to temporal resolution reduction. Also, videos with temporal distortions, like packet loss, are very sensitive to temporal resolution reduction.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acronyms

BBC British Brodcasting Corporation. 6

DCT Discrete Cosine Transform. 4, 6, 34, 35

DFT Discrete Fourier Transform. 24

DMOS Difference Mean Opinion Scores. 40, 42, 47

FPS Frames Per Second. xii, 35, 36, 40, 41, 43, 56, 57, 60

GMSD Gradient Magnitude Similarity Deviation. 17, 19, 20, 44–46, 52, 53, 57, 58, 63, 64

GOF Group of Video Frames. 26

GSM Gaussian Scale Mixture. 22

HVS Human Visual System. 4, 15–17

INLSA Iterated Nested Least-Squares Algorithm. 47

ITU International Telecommunication Union. 15

JND Just Noticeable Differences. 16

MOAVI Monitoring Of Audiovisual Quality by Key Indicators. 33

MOS Mean Opinion Scores. 1, 41, 43, 47

MSE Mean Squared Error. 15, 17, 25

PSNR Peak Signal-to-Noise Ratio. 1, 6, 15, 17, 18, 44–46, 49, 52, 53, 55, 57, 58, 63, 64

RMS Root Mean Squared. 25, 29, 31

RRED Reduced Reference Entropic Differencing. 21, 22

SA Spatial Activity. 31, 33

SAMVIQ Subjective Assessment Methodology for Video Quality. 41

SCC Spearman Correlation Coeficient. xi, 43, 44, 52–56, 72

SRRED Spatial Reduced Reference Entropic Differences. 21–23

SSEQ Spatial–Spectral Entropy-based Quality. 34–36, 54, 55, 64

SSIM Structural Similarity Index. 1, 16–19, 44–46, 49, 52, 53, 57, 58, 63, 64

SSTS-GMSD Spatial and Spatial-Temporal Slices Gradient Magnitude Similarity Deviation. 17, 20, 27, 39, 45, 46, 48, 49, 52, 53, 57, 63

STRRED Spatial-Temporal Reduced Reference Entropy Difference. 17, 21, 23, 45, 46, 48, 49, 51–58, 63, 64

STS Spatial-Temporal Slices. 17, 20

SVM Support Vector Machine. 34, 35, 53

TRRED Temporal Reduced Reference Entropic Differences. 22, 23

ViS3 Video Quality Assessment via Analysis of Spatial and Spatial-Temporal Slices. 17, 23, 26, 45, 46, 48, 49, 52–58, 63, 64

VQEG Video Quality Experts Group. 38

1 INTRODUCTION

In recent years there has been an increasing demand for video-based services. According to Cisco in 2019, 80 percent of all consumer Internet traffic will be from video applications [5]. As consume increases, the demand for better user experiences also increases. Many factors contribute to the quality of the user experience, like the display in which the video is playing, the ambient light, the user interest, and the video quality. Video quality is the quality of the video signal as perceived by the user.

There are basically two ways of assessing the quality of a video: subjectively and objectively. Subjective video quality assessment methods estimate the quality of a video by performing a series of pycho-physical experiments. In these experiments, subjects watch a series of videos and give a quality score to each the video. The average of these scores is the Mean Opinion Scores (MOS), which is a subjective quality estimate. This type of video quality assessment method is known to be the most reliable method [6]. However, psycho-physical experiments are expensive, very difficult to replicate, and require a minimum number of subjects taken from a diverse pool of people to provide reliable results. Most of the times, objective video quality assessment methodologies are used. Objective methods are basically algorithms that automatically compute a quality estimate score for the video. These methods are faster and cheaper than the subjective methods, but they are less precise.

In the early days, image quality assessment methods were adapted to measure the video quality. In other words, image quality assessment methods, like Peak Signal-to-Noise Ratio (PSNR), were computed for each frame and averaged to provide the video quality predicted score [7]. However, using image quality assessment methods has limitations. The main issue is that these methods do not account for temporal distortions. To solve this problem, some research add temporal information to image quality assessment methods to improve their prediction accuracy performances. For example, Wang *et al.* propose adding a motion estimation stage to the Structural Similarity Index (SSIM) [1], which adds more weight to slow moving regions [8]. Vu *et al.* adapt the Most-Apparent-Distortion (MAD), an image quality assessment method, to video by using spatial-temporal slices and optical-flow [9].

Nowadays, most of the proposed video quality assessment methods use spatial and temporal features to evaluate video quality. For example, Pinson *et al.* extracted several features from the distorted and reference video to estimate the video quality [10]. Some of the features extracted by this method include spatial activity, temporal activity, and color-based differences. Seshadrinathan used a 3D Gabor analysis to extract spatial, temporal and spatial-temporal features from the distorted video and compared them with the same features extracted from the reference video [11].

Although video quality assessment methods are becoming more precise, there are still chal-

lenges in this area. Chandler and Wang discuss some of these challenges related to image/video quality assessment methods [12, 13]:

- ***How to deal with multiple distortions*, many video quality assessment methods have a very good performance when the content evaluated has the same distortion. Because, only the level of distortion is modified from one video to another. However, different video distortions affect the video in different ways, and sometimes the video quality methods cannot interpreted which distortion is less visible.**
- ***How to deal with image enhancements*, some video/image post-processing can improve the perceptual quality of the video/image, but the video quality assessment can interpreted this processing as a distortion.**
- ***How to make the image/video quality assessment scores easy-to-use and easy-to-understand*, image/video quality assessment methods are becoming more complex over the years. Many methods have great correlation with subjective scores, however they are complex algorithms that are hard to understand and use.**
- ***How to estimates video quality of High Dynamic Range (HDR) content*, this type of content are becoming more common and many video quality assessment methods are not ready to evaluate this type of content.**

Finally, one big challenge**, that is mention in both articles**, is reducing the runtime performance. Some of the current video quality methods are so computationally complex that it is impossible to use them in real time scenarios or any practical application. For example, the average runtime of a recent developed method, the ViS3, is more than 600 seconds for videos 10 seconds long [12]. Some video quality methods have options to reduce the runtime. For example, Wang's video quality method has an option to reduce the number of frames used by the algorithm to estimate the video quality [8]. The VQM reduces the spatial and temporal resolution of the video to improve the runtime performance [10]. And MOVIE analyzes the video in an 8-by-8 frame interval [11]. Although these methods offer options to improve the runtime performance, they do not consider that different video distortions are more or less sensitive to reductions of spatial or temporal resolution.

Our main objective is to create a framework to improve the runtime of video quality methods without affecting their prediction accuracy performance. One of the simplest method to improve runtime performance is to reduce the video spatial resolution. **However, when the video resolution is reduced, the video quality is alter and the accuracy performance of video quality assessment methods can be decrease. So, we made several tests using six video quality assessment methods and analyze their accuracy performance, when reducing the video spatial resolution.** We analyze the effects of reducing the spatial resolution and also try to understand which video distortions are more or less sensitive to these reductions. We propose a four step framework. The first step consists of identifying the videos which are more sensitive to spatial

resolution reduction. The second step is to reduce the video resolution according to its distortion. The third step consist of measures the video quality using an objective video quality assessment method. Finally, the fourth step is to adjust the predicted score from video with different resolutions to the same scale.

In this work, we design two video distortion classifiers to perform the first step of the proposed framework. These classifiers identify videos more sensitive to spatial resolution reduction, so that the spatial resolution can be adjusted accordingly. We tested the proposed framework with six objective quality assessment methods. We notice an improvement in runtime performance for the six methods and, surprisingly an improvement in the prediction accuracy performance.

Also, we perform tests to analyze the effects on the video quality assessment methods accuracy performance, when the temporal resolution is reduced. Reducing the temporal resolution is another simple and fast way to reduce the runtime performance of video quality assessment methods.

We divide this work into six chapters. In Chapter 2, we explain how digital video is transmitted and detail the most common video distortions. In Chapter 3, we describe the video quality assessment methods used in this work. In Chapter 4, we detail the proposed framework and video classifiers. In Chapter 5, we present the results of this work. Finally, in Chapter 6, we present the conclusion and future works.

2 DIGITAL VIDEO SYSTEM

Before a video reaches the final user, it has to go through some processing stages. Figure 2.1 shows an overview of a common digital video transmission system. First, the original video is compressed to reduce its size and, consequently, increase the transmission speed. The encoding stage includes the compression and the channel encoding. The result of the video compression stage is a bitstream, in which the coded data is divided into packets for transmission. These packets are encoded according to the transmission protocol requirements. Then, the video is transmitted through a channel that can be dynamic or static, packet-switched or circuit-switched. When the video data reaches the receiver, it is decoded and decompressed in the decoding stage [14].

Figure 2.1: Overview of Video Transmission Diagram.

During the transmission process, the video can be distorted in different ways. The compression of the video, usually, is a lossy compression, meaning that some of the video data is discarded and the received video cannot be perfectly reconstructed [15]. Depending on the type of compression and the compression rate, some visible artifacts can be present in the video. Also, during the transmission through the physical channel, some of its packets can be lost, what may also affect the video quality. After the video is received and decompressed, it can further be processed to a better visualization on the device that it is been displayed. For example, if the video is in a lower resolution than the display, the video can be upscaled to the display resolution. Finally, the video can be corrupted by noise in many ways, during the transmission or processing stages [16]. **Next, we describe the most relevant video distortions and how they are generated in a modern video transmission scenarios.**

2.1 VIDEO PROCESSING AND VIDEO DISTORTIONS

Video distortions are perceptible differences between the degraded video and the corresponding reference video. These distortions are the result of one or more errors introduced in the videos during compression, post processing or transmission errors. Because compression algorithms are very complex and have different stages, videos distorted by compression usually are affected by more than one artifacts (a visible distortion). The post processing of a video usually has the goal of attenuating artifacts. Finally, transmission errors occurs because of channel errors, like losses and congestion problems.

2.1.1 Video Compression Algorithms

Data compression techniques have the goal of representing the data in a more compact form [15]. Videos are compressed to reduce storage and transmission costs. There are lossless and lossy compression algorithms. Although lossless compression algorithm can reduce the video size and recover it without any errors, the compression ratio that can be achieved with lossless compression algorithm is limited [17]. Fortunally, since video sequences are usually very redundant, lossy compression algorithms can achieve great compression rates with a good video quality. The most popular lossy compression algorithms are MJPEG, MPEG-2, DIRAC-Wavelet, H.264 and HEVC. Although lossy compression algorithms are more efficient in terms of compression rate, they may introduce perceivable artifacts into the video.

The MJPEG compression algorithm is based on the JPEG image compression. JPEG makes use of the Discrete Cosine Transform (DCT). First, the image is divided into 8x8 blocks and the DCT of each block is computed. The DCT coefficients are quantized, removing most of the high frequency coefficients. The quantized coefficients are rearranged and coded using a Huffman coding algorithm [18]. In MJPEG each video frame is compressed as a JPEG image. Although JPEG is very good at compressing images, using it to compress individuals video frames is not a very good strategy. Video sequences have temporal redundancies that can be explored to allows for a more efficient compression of video sequences. So, videos compressed by MJPEG have a lower video quality than videos compressed by other compression standards, like MPEG-2 and H.264, which take into account temporal redundancies.

Figure 2.2 shows an example of video frame compressed by MJPEG, which is affected by blocking and blurring artifacts. The blocking artifact is a consequence of the quantization of the DCT coefficients. Since the coefficients in each block are processed independently, regardless of the spatial correlation between them, intensity differences between neighboring blocks cause the edges/borders that characterize blockiness. The Human Visual System (HVS) can easily detect blocking artifacts because of equidistant distribution of the blocks in the frame [19]. Blurring artifacts, on the other hand are caused by any type of operation that discards high frequencies DCT coefficients [19]. In image processing, the high frequencies coefficients are associated with edges and abrupt changes in intensities [20]. Therefore, blurred images have less sharp edges and borders.

MPEG-2 compression algorithm is a block-based motion-compensated video coding algorithm. It is similiar to JPEG, but it uses motion estimation techniques to reduce temporal redundancy. Since it also quantizes the DCT block coefficients, discarding most of the high frequencies components, MPEG-2 may also introduce blockiness and blurring artifacts in the compressed video. The motion estimation techniques search blocks in previous frames that are similar to blocks in the current frame [21]. This way, only the coordinates of the similar blocks and the difference between the similar blocks and the current blocks need to be encoded and sent to the receiver. Videos compressed by MPEG-2 may also be affected by mosquito noise, as a consequence of the motion estimation. This artifact occurs when the predicted block contains only part

(a) Reference Frame

(b) MJPEG

Figure 2.2: Comparison between (a) the reference frame and (b) the frame compressed using a MJPEG algorithm.

of a moving object. Mosquito noise is more visible in smooth texture regions around egdes [14]. Figure 2.3 shows a video frame compressed by MPEG-2 and the corresponding reference frame.

H.264 compression algorithm is also a block-based motion-compensated video coding algorithm [22]. In H.264, the video is first divided into macroblocks (16x16 for the luminance component and 8x8 for the chroma components), and, then, the DCT coefficients of each macroblock are calculated and quantized. Next, a motion estimation algorithm is used to obtain inter-frame predictions. Different from previous compression algorithms, each macroblock can be divided into sub-macroblocks down to 4x4, which improves the quality of the prediction, specially for fast motion regions. One of the major features of H.264 to improve video quality is the deblocking filter to reduce the blocking artifacts. Blocking artifacts are common in other block-based compression algorithms. Koh *et al.* show that the blocking is the most annoying artifact present in videos compressed by MPEG-2 [23]. The use of deblocking filter improves the quality of the video (by 9% in PSNR) [24]. Figure 2.4 shows a video frame compressed by H.264 compression algorithm and the corresponding reference frame. We can notice that the blocking artifact is less visible, comparing with the frames distorted by MJPEG and MPEG-2. However, video frames compressed with H.264 still present blurring artifacts.

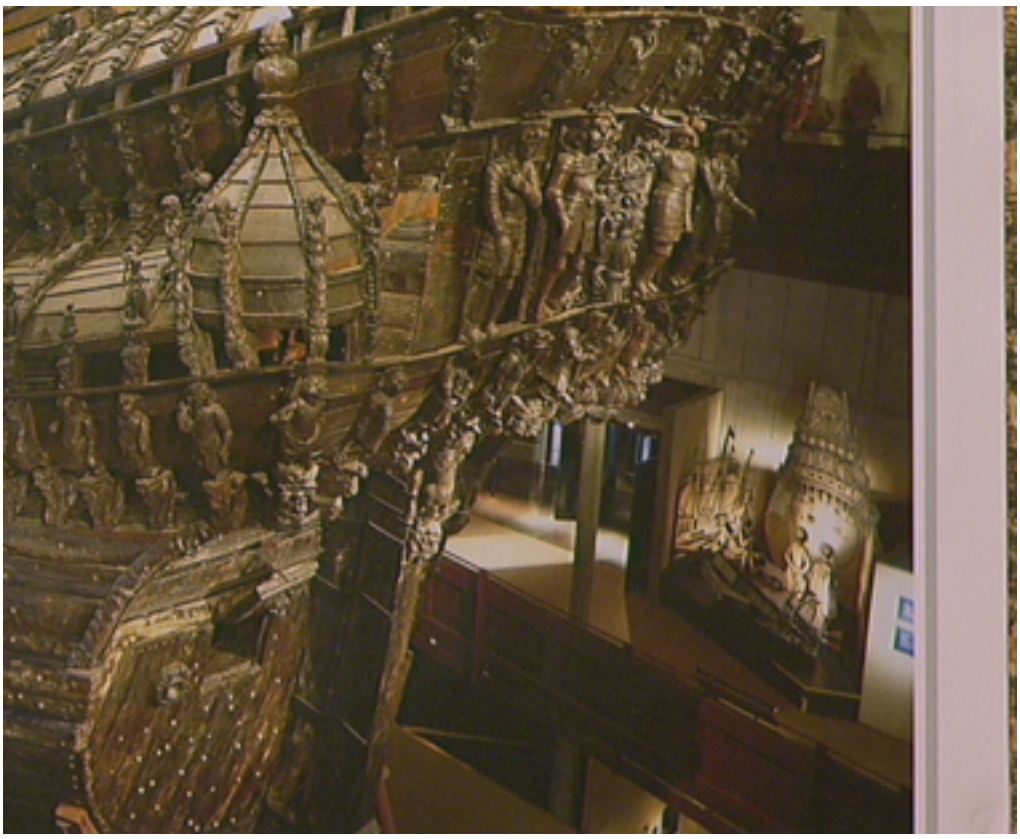
DIRAC-Wavelet compression algorithm was developed by the British Brodcasting Corporation (BBC) [25]. The major difference between the DIRAC-Wavelet algorithm and other compression algorithms, like MPEG-2 and H.264, is that it uses wavelets instead of DCT, which reduces the blockiness artifacts. But, the quantization of wavelet coefficients leads to ringing and blurring artifacts. Ringing artifacts manifest themself as wave-like structures in edges regions. Figures 2.5 shows an example of a video frame distorted by a DIRAC-Wavelet compression. Notice that ringing artifacts are present in areas around high contrast edges.

H.265 or HEVC compression algorithm is an improved version of the H.264 compression. One of the changes introduced by H.265 is the possibility of partitioning the video frame using blocks of different sizes, which varies from 64x64 down to 8x8 [26]. The performance of H.265 is 50% better than the performance of H.264. For the same quality level, the bitrate obtained with H.265 is half the bitrate obtained with H.264 [27]. Figure 2.6 shows a video frame distorted by an H.265 compression. Videos compressed with H.265 at low bitrates contain mostly blurring artifacts.

2.1.2 Upscaling Processing

Sometimes the display where the video is presented has a higher resolution than the video itself. Algorithms of interpolation are then used to upscale the video resolution to the display resolution. This type of processing is commonly used in streaming services applications, because video resolutions vary according to the available bandwidth [28].

Interpolation algorithms used in upscaling consider the neighboring pixels to estimate a pixel value. **There are different interpolation algorithms. Nearest neighbor, biliniar and bicubic**



(a) Reference Frame

(b) MPEG-2

Figure 2.3: Comparison between (a) the reference frame and (b) the frame compressed using a MPEG-2 compression algorithm.

(a) Reference Frame

(b) H.264

Figure 2.4: Comparison between (a) the reference frame and (b) the frame compressed using an H.264 compression algorithm.

(a) Reference Frame

(b) DIRAC-Wavelet

Figure 2.5: Comparison between (a) the reference frame and (b) the frame compressed using a DIRAC-Wavelet compression algorithm.

(a) Reference Frame

(b) HEVC

Figure 2.6: Comparison between (a) the reference frame and (b) the frame compressed using an H.265 compression algorithm.

(a) Reference Frame

(b) H.264

(c) H.264 with Upscaling

Figure 2.7: Comparison between (a) the reference frame and (b) the frame compressed by an H.264 algorithm and (c) the frame compressed by an H.264 algorithm with Upscaling.

(a) Reference Frame

(b) White Noise

Figure 2.8: Comparison between (a) the reference frame and (b) the frame distorted by white noise.

(a) Reference Frame

(b) Packet Loss

Figure 2.9: Comparison between (a) the reference frame and (b) the frame distorted by packet loss artifacts.

the most common interpolation algorithms [29]. The nearest neighbor algorithm is the most simple interpolation algorithm, and generates image with less quality than the other methods. The biliniar and bicubic generates images with better quality, because they use more information to predicted the new pixel value. However, these interpolation algorithms are low-pass filter and smooths the edge regions in the frames. Figure 2.7 shows an example

of video frame distorted by upscaling. In this example, the video was compressed with H.264 and then upscaled to a higher resolution, using a bilinear interpolation algorithm. Notice that the upscaled video frame is more blurred than the reference video frame.

2.1.3 Video Transmission

Video can be transmitted either in analog or digital channels. In analog transmissions, one of the most common artifacts is noise, in particular white noise that is a random signal that occupies the whole spectrum. Noise can also be added in video sequences during the acquisition and processing [16]. Figure 2.8 shows a video frame distorted by white noise. Notice that in this case the noise is an additive distortion, which affects the whole frame.

In digital transmission, before the compressed video is transmitted over a channel (e.g. a wired or wireless transmission channel), it is divided into packets. Sometimes these packets are discarded due to the traffic in the network. The impact of a packet loss in video quality depends on the content of the video and the channel loss distribution [30]. Packet loss is characterized by the presence of erroneously decoded blocks in the decoded video. **In most video compression algorithms, frames that are coded without any reference to previous frames are send periodically. These frames are called I-Frames, and they are used as a reference to decode other frames.** If the packet loss occurs in these I-Frames, the distortions are more severe and may affect several frames [31]. Figure 2.9 shows a video frame distorted by packet loss. Notice that the packet loss is a local distortion, i.e. it is spatially and temporally concentrated.

2.2 VIDEO DATABASE

Video quality database are a collection of videos and their distorted versions. Each distorted video has a subjective score that represents the quality of the video. These databases labeled the distorted videos by its distortion, as shown in the previously section. In this work, we use four databases: LIVE, CSIQ, IVP and MCL-V. Each database has unique specifications and they are describe in the next subsections

2.2.1 LIVE Video Quality Assessment Database

The LIVE Video Quality Assessment Database is one of the most popular video quality database [32, 31]. It was released in 2010 by the Laboratory for Image & Video Engineering (LIVE) of the University of Texas. This database has 10 reference videos and, for each reference video, there are 15 test sequence with four types of distortions: MPEG-2 compression (4 test videos per reference), H.264 compression (4 test videos per reference), simulated transmission of H.264 compressed bitstreams through error-prone IP networks (3 test videos per reference) and through error-prone wireless networks (4 test videos per reference). Figure 2.10 shows sample

frames from each reference video in the LIVE database.

The videos are provided in uncompressed YUV420 format at a 768x432 resolution. Nine of the ten reference videos are 10 seconds long and one video is 8.68 seconds long. Seven out of the ten reference videos have 25 FPS and the other three videos have 50 FPS.

For the subjective testing, a single stimulus methodology was adopted. In this methodology, the subject watches a single video and then gives a quality score to this video using a continuous scale. All test videos were evaluated by 38 subjects. Subjects evaluated reference and the impaired videos. The Difference Mean Opinion Scores (DMOS) between each impaired video and the corresponding reference was computed to provide a subjective quality score.

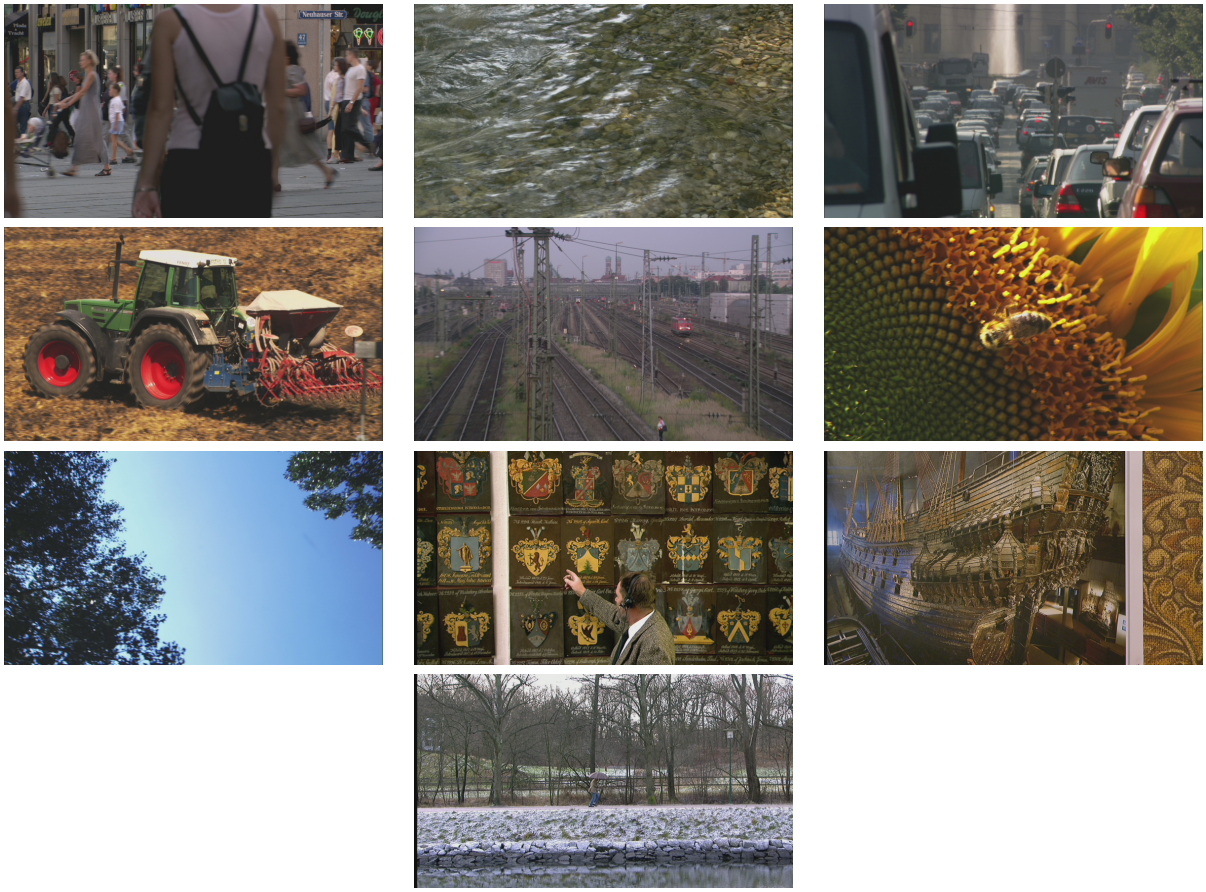


Figure 2.10: Frames of all videos in LIVE Database

2.2.2 CSIQ Video Quality Database

The CSIQ Video Quality Database was released in 2013 by Laboratory of Computational Perception & Image Quality at Oklahoma State University [4]. This database contains 12 reference video sequences and 216 distorted video sequences. All videos were provided in uncompressed YUV420 format with a spatial resolution of 832x480. Figure 2.11 shows sample frames from each reference video in the CSIQ database.

The distorted video sequences were generated using 6 types of distortion: H.264 compression, H.265 compression, MJPEG compression, wavelet-based compression using the Snow codec, H.264 videos subjected to simulated wireless transmission loss, and additive white noise. To generate all test videos, each reference video was distorted with all distortions in three different levels.

The Subjective Assessment Methodology for Video Quality (SAMVIQ) was used in the subjective tests [33]. SAMVIQ is a multi-stimulus experimental methodology. Subjects are presented with the reference and distorted sequences from the same content. They evaluate each sequence in any order. They can even change the quality score given to a sequence previously seen. All videos were evaluated by 35 subjects. The MOS for each video sequence was released with the database.

Figure 2.11: Frames of all videos in CSIQ Database

2.2.3 IVP Subjective Quality Video Database

The IVP Subjective Quality Video Database was released in 2011 by the Image and Video Processing Laboratory at The Chinese University of Hong Kong [34]. This database contains 10 reference video sequences and 128 distorted video sequences. The videos were provided in uncompressed YUV420 format with resolution of 1920x1088. All videos have 25 FPS. Seven

of the reference video sequences are 10 seconds long, two are 11.2 seconds long and one is 8.96 seconds long. Figure 2.12 shows sample frames from each reference video in the IVPL database.

The distorted video sequences were generated using 4 types of distortion: MPEG-2 compression (3 test videos per reference), wavelet compression (3 test videos per reference), H.264 compression (4 test videos per reference) and packet loss cause by H.264 streaming through IP networks (4 test videos per reference). The packet loss distortions are presented only for seven (out of ten) reference video sequences.

All videos were evaluated by 42 subjects. The single stimulus methodology was used in the subjective experiment. DMOS was provided as the subjective quality measure.

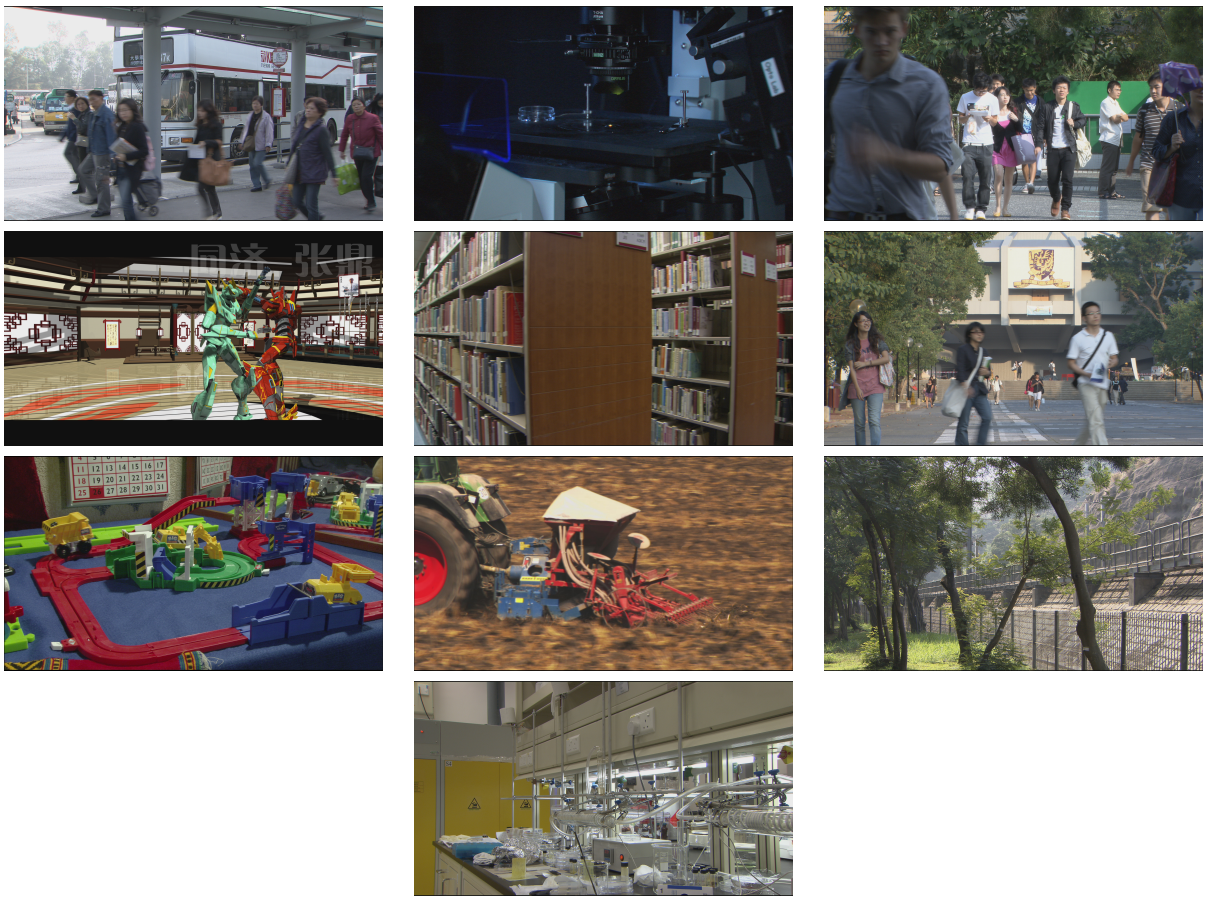


Figure 2.12: Frames of all videos in IVP Database

2.2.4 MCL-V Database

The MCL-V Database was released in 2014 by the USCMediaCommLab at the University of Southern California [28]. This database contains 12 reference videos sequences and 96 distorted video sequences. The distorted video sequences were generated using 2 types of distortion: H.264 Compression (4 distorted video sequences per reference) and H.264 Compression follow by scaling (4 distorted video sequences per reference). Figure 2.13 shows frames from each reference video in the MCL database.

The video sequences were provided in uncompressed YUV420 format with a spatial resolution of 1920x1080. Two of the reference video sequences have 24 FPS, six of them have 25 FPS and the remaining sequences have 30 FPS. The videos sequences with 25 and 30 FPS are 6 seconds long. And one of the video sequences with 24 FPS is 5.5 seconds long and the other video sequence is 5 seconds long.

The videos were evaluated by 45 subjects. A pairwise comparison experimental methodology was used in the subjective test. This methodology is recommended by ITU [35], and consists in simultaneously presenting two video sequences from the same source, but with different quality levels. Subjects are asked which of the video sequences has a better quality. The database provides the MOS of each test sequence.

Figure 2.13: Frames of all videos in MCL-V Database

3 VIDEO QUALITY ASSESSMENT METHODS

There are two ways to evaluate the video quality. The first way is the subjective quality assessment method. In these methods, psychophysical experiments are performed in which subjects evaluate video quality using a scale directly comparing video sequences. To perform a subjective test properly, experimenters follow recommendations established by the International Telecommunication Union (ITU) [36]. Subjective methods are known to provide the most reliable results for video quality, however, they are very expensive and time-consuming.

The second way is the objective assessment method. This method uses mathematical models, which can be implemented in hardware or software, to evaluate video quality. Depending on the information available, objective methods can be categorize into three groups:

- Full-Reference: the original and the test video are available to the method at the measurement point.
- Reduced-Reference: some information about the original video and the test video are available to the method at the measurement point.
- No-Reference: only the test video is available to the method at the measurement point.

Full-Reference methods are often used in non-real-time scenarios, because these methods require information about the reference video, which is frequently not available in real-time applications. Codec comparison and optimization are examples of scenarios in which these methods can be used. On the other hand, reduced-reference and no-reference methods are more appropriate for real-time scenarios. Nevertheless, full-reference methods are more precise than no-reference and reduced-reference methods.

Objective quality assessment methodologies can also be classified as data metrics or picture metrics [37]. Data metrics measure only the fidelity of the video signal, i.e. how similiar are the test and reference video. Data metrics do not consider the content of the video or how degradations are perceived by human viewers. These metrics, often, are simple and fast. Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) are data methods. One of the problems of data metrics is that they do not consider the type of distortion present in the video. Figure 3.1 shows five images with different types of distortions and different quality levels, which have same MSE score. As can be seen in these images, although the HVS perceived these distortions differently, MSE does not capture these differences. Another problem, with data metrics is that they do not take into consideration the content of the video. And according to Wang, distortions may be less visible in texture areas of the video frame or when the speed of motion is large [8]. Nevertheless, it is worth pointing out that data metrics can be used in other contexts, when we simply want to measure the differences between two video signals.

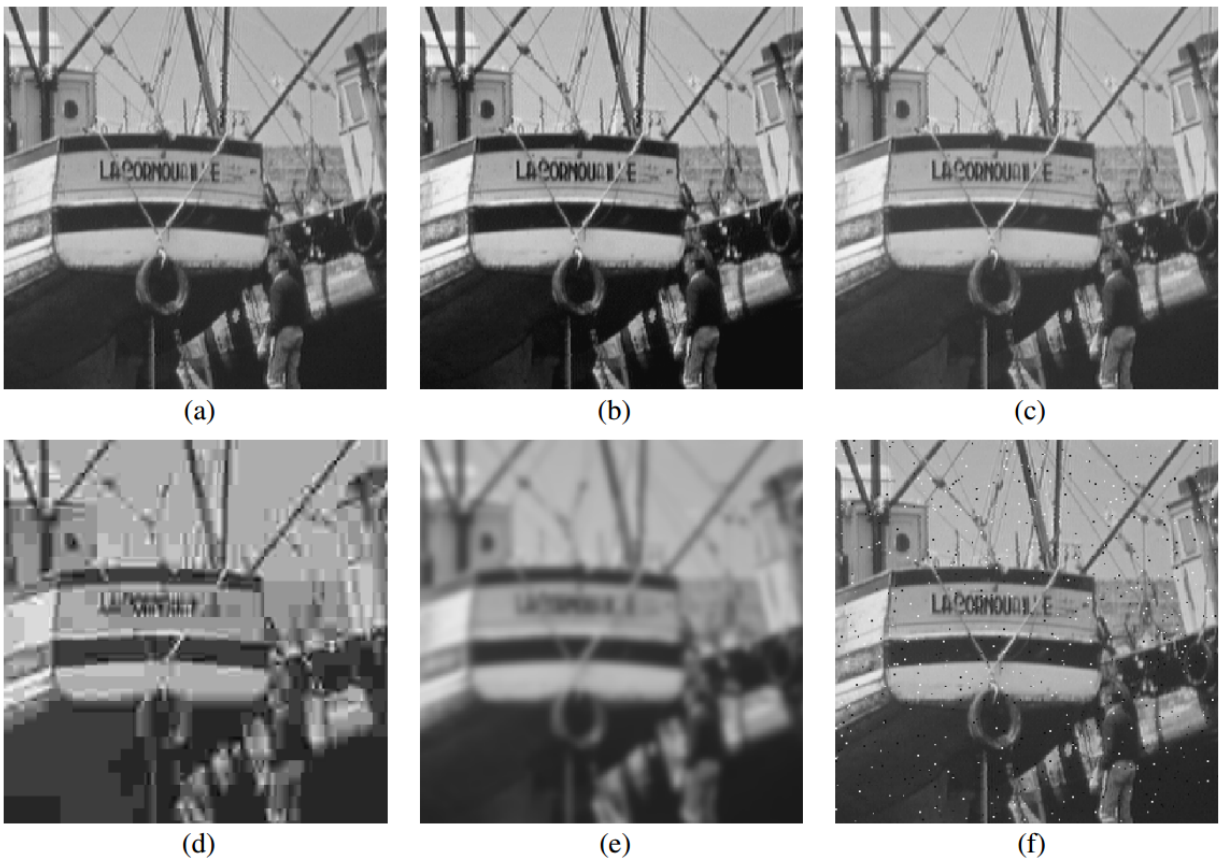


Figure 3.1: Comparison between the images with the same MSE score, but with different disotrtions. (a) Orignal Image. (b) Contrast-stretched image. (c) Mean-shifted image. (d) JPEG compressed image. (e) Blurred image. (f) Salt-pepper impulsive noise contaminated image. Orignal From [1].

In the last decades, researchers started developing pictures metrics, which are objective quality assessment methods that take into consideration the intrinsic characteristics of a visual signal. In other words, picture metrics analyze the visual data as visual information, considering the effects of video distortions and content on the perceived quality. Picture metrics can be designed either using a vision modeling approach or an engineering approach. The vision modeling approach implements models of components of the HVS, like color perception, contrast sensitivity, and pattern masking. One example of a video quality assessment method that uses a vision modeling approach is the Sarnoff Just Noticeable Differences (JND) method, which is based on the chromatic and luminance differences that can be perceived by human viewers, i.e. the just noticeable perceived difference [38]. Another example of a video quality assessment method that uses visual models is the Perceptual Distortion Metric (PDM), which is based on some aspects of the human vision like perceptual color space, multi-channel representation of spatial and temporal mechanisms, contrast sensitivity, and pattern masking [39].

The engineering approach is based on the analysis of certain features of the video, like the spatial luminance gradient, the image structure, or specifics artifacts (like blur, noise, etc). Some examples of video quality assessment methods that adopt an engineering approach are: the SSIM, an image quality assessment method, that measures the mean, variance and covariance of patches

of the image [1]; the Gradient Magnitude Similarity Deviation (GMSD), an image quality assessment method that analyzes the spatial luminance gradient [40]; and the Spatial and Spatial-Temporal Slices Gradient Magnitude Similarity Deviation (SSTS-GMSD) that is an extended version of the GMSD, that analyzes the gradient of the Spatial-Temporal Slices (STS) of the video [2].

On the other hand, there are methods that combine both approaches. One example is the Video Quality Assessment via Analysis of Spatial and Spatial-Temporal Slices (ViS3), which includes a data metric and a picture metric in its model [4]. First, a spatial distortion map is computed using the Most Apparent Distortion algorithm (MAD), which takes into account on how the HVS perceives video distortions [41]. Then, the spatial-temporal dissimilarity between the reference and distorted videos is computed using spatial-temporal correlation and spatial-temporal responses.

Recently, the use of machine learning algorithms for video quality assessment become extremely popular, specially for the no-reference quality assessment methods scenario. Most machine learning based quality assessment methods extract features from the videos and use machine learning techniques, like support vector regression or neural networks, to predicted the video quality scores. Xu *et al.* propose a no-reference video quality assessment method that uses an unsupervised feature learning approach that trains a support vector regression to predict video quality [42]. Saad *et al.* use a spatio-temporal natural scene statistics model and a motion model to train a support vector regression algorithm [43].

Although prediction accuracy performance of video quality assessment methods are improving, these methods are becoming more complex, requiring large running times. Both ViS3 and Spatial-Temporal Reduced Reference Entropy Difference (STRRED) take several minutes to estimate the quality of a video 10 seconds long [4, 44]. MOVIE takes hours to estimate the quality of a video 10 seconds long [11], because the algorithm uses a 3D Gabor filter. Improving the runtime of video quality assessments methods is still an open challenge [12, 13].

In this work, we study the performance of full-reference and reduced-reference video quality assessment methods. Among the assessment methods studied in this work are the full-reference assessment methods: ViS3, SSTS-GMSD, GMSD, SSIM and PSNR; and the reduced-reference assessment method STRRED. These methods are detailed in the next sections.

3.1 PSNR

The Peak Signal-to-Noise Ratio (PSNR) is a simple full-reference assessment method that is based on MSE, that is the mean of the squared difference of to images, as given by the following equation:

$$MSE = \frac{\sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (I_{ref}(x,y) - I_{dist}(x,y))^2}{M \cdot N}, \tag{3.1}$$

where $I_{ref}(x,y)$ is a pixel from the reference image, $I_{dist}(x,y)$ is a pixel from the distorted or test image and MxN is the spatial dimension of these images. In this work, we consider only the luminance component of the image to calculate the MSE. The PSNR is calculated using the following equation:

$$PSNR = 10 \cdot log_{10} \frac{MAX_I^2}{MSE}, \tag{3.2}$$

where MAX_I is the maximum possible value of the image (generally $MAX_I = 255$). To use PSNR as a video quality method we calculate the PSNR for each frame of the video and take the average of the frame results to obtain the video quality score.

3.2 SSIM

The Structural Similarity (SSIM) is a full reference image quality assessment method developed by Wang [1]. SSIM compares three features of the reference and distorted image. These features are luminance, contrast and structure, as shown in Figure 3.2.

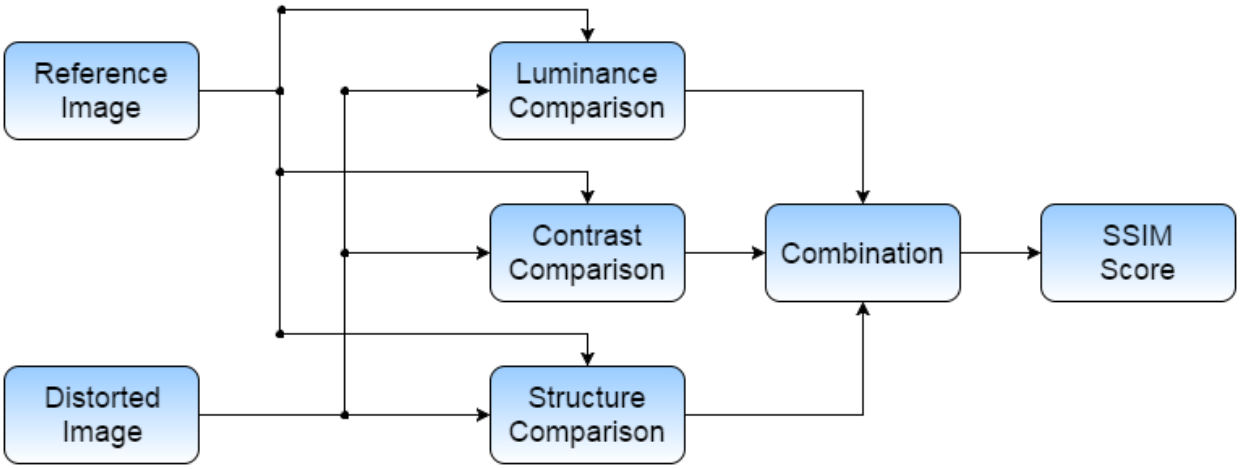


Figure 3.2: Diagram of SSIM Algorithm.

SSIM performs its analysis on the luminance component of the image. To compute the SSIM 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:

$$l(x,y) = \frac{2\mu_x\mu_y + C}{\mu_x^2 + \mu_y^2 + C}, \tag{3.3}$$

where μ_x and μ_y 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:

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{3.4}$$

where σ_x and σ_y 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:

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \tag{3.5}$$

where σ_x and σ_y are the standard deviation intensity of the original and distorted block, σ_{xy} 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:

$$SSIM(x,y) = l(x,y)^\alpha \cdot c(x,y)^\beta \cdot l(x,y)^\gamma, \tag{3.6}$$

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.

3.3 GMSD

The Gradient Magnitude Similarity Deviation (GMSD) is a full reference image quality assessment method developed by Xue *et al.* [40]. It is based on the gradient differences between reference and distorted (test) gray-scaled images. The gradient magnitude is computed using Prewitt filters for horizontal and vertical directions, as given by the following expressions:

$$h_x = \begin{vmatrix} 1/3 & 0 & -1/3 \\ 1/3 & 0 & -1/3 \\ 1/3 & 0 & -1/3 \end{vmatrix} \tag{3.7}$$

$$h_y = \begin{vmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 \\ -1/3 & -1/3 & -1/3 \end{vmatrix} \tag{3.8}$$

To compute the gradient magnitude image, we convolve the reference and distorted (test) images with the Prewitt filters, as shown in the following equations:

$$m_r = \sqrt{(I_{ref} * h_x)^2 + (I_{ref} * h_y)^2}, \tag{3.9}$$

$$m_d = \sqrt{(I_{dist} * h_x)^2 + (I_{dist} * h_y)^2}, \tag{3.10}$$

where I_{ref} is the reference image, I_{dist} is the distorted image, m_r is the gradient magnitude of the reference image, m_d is the gradient magnitude of the distorted image, and $*$ represents the convolution operation. Finally, we compare the gradient magnitude of the reference and distorted images using the following equation:

$$GMS_{map} = \frac{2m_r m_d + c}{m_r^2 + m_d^2 + c}, \tag{3.11}$$

where c is a small constant necessary to avoid instability. The GMSD final score is obtained by calculating the standard deviation of the GMS_{map}. To use GMSD as a video quality method we calculate the GMSD for each video frame and average these to obtain the video quality score.

3.4 SSTS-GMSD

The Spatial and Spatial-Temporal Slices Gradient Magnitude Similarity Deviation (SSTS-GMSD) is a full-reference video quality assessment method based on GMSD, which was developed by Yan *et al.* [2]. It uses Spatial-Temporal Slices (STS) of the video to provide temporal information to the GMSD. A video can be represented as a 3-D array, $F(x, y, t)$, where the x and y are spatial coordinates and the t is a temporal coordinate. The STS are represented by the perpendicular dimension to x and y. Figure 3.3, depicts the STS. There are two types of STS: the vertical STS, represented by x and t coordinates, and the horizontal STS represented by y and t coordinates.

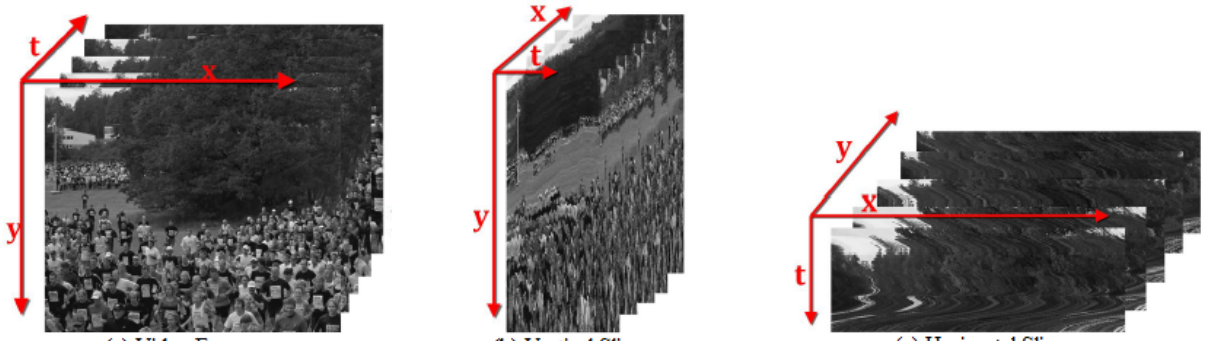


Figure 3.3: Example of Spatial Temporal Slices. Originals taken from [2].

The SSTS-GMSD algorithm computes the GMSD of the video frames (*x,y*), the horizontal STS (*x,t*) and the vertical STS (*y,t*), as shown in Figure 3.4. Then, the GMSD scores are sorted in descending order and the average value of the first 20% GMSD scores is computed, for each of these dimensions. This creates the Spatial GMSD index (S-GMSD), the Horizontal-Slice index

(H-GMSD), and the Vertical-Slice GMSD index (V-GMSD). The following equation shows how these indices are combined into a single index that represents the quality of the video:

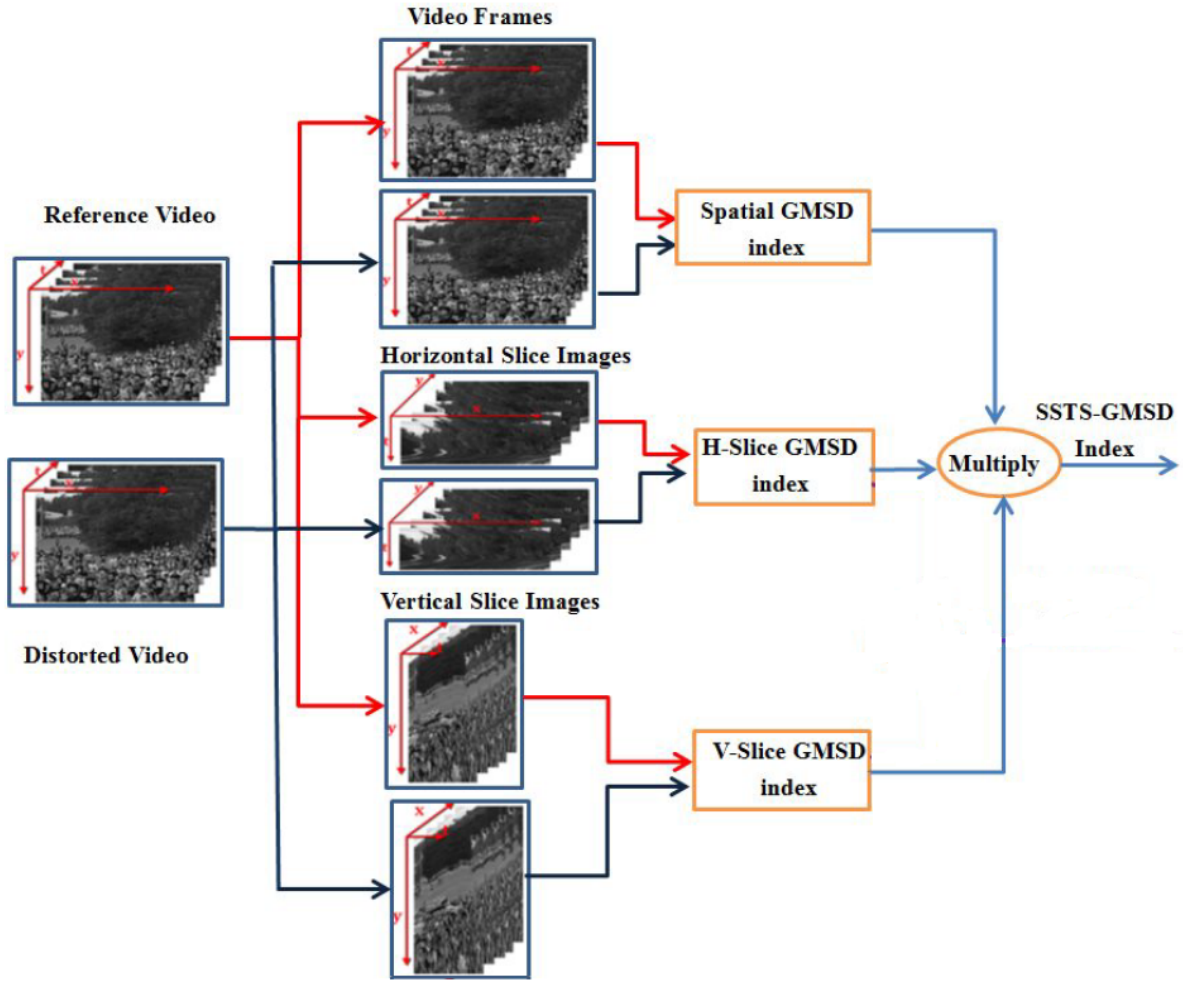


Figure 3.4: Diagram of SSTS-GMSD. Original from [2].

$$SSTS\text{-}GMSD = \sqrt{S\text{-}GMSD \cdot V\text{-}GMSD \cdot H\text{-}GMSD}. \tag{3.12}$$

3.5 STRRED

The Spatial-Temporal Reduced Reference Entropy Difference (STRRED) is a reduced reference video quality method developed by Soundararajan and Bovik [44]. It is based on a previously work, the Reduced Reference Entropic Differencing (RRED) [45]. RRED is a reduced-reference image quality assessment method based on the differences between the entropies of wavelet coefficients of the reference and distorted images. The STRRED estimates video quality based on spatial and temporal distortions. The spatial distortions are measured using the Spatial Reduced Reference Entropic Differences (SRRED), which is based on the frame's RRED index with the ad-

dition of a motion weight, make it more sensitive to distortions that occur in slow motion regions. Temporal distortions are measured using the Temporal Reduced Reference Entropic Differences (TRRED), which is based on the adjacent frames difference RRED index.

The SRRED algorithm calculates the entropic differences of the wavelet coefficients between reference and distorted frames. First, the wavelet coefficients are calculated using a steerable pyramid decomposition with multiple scales and orientations [3]. **Figure 3.5 shows an example of a steerable pyramid decomposition in three scales and three orientations. Each image in the steerable pyramid decomposition is considered a sub-band.**

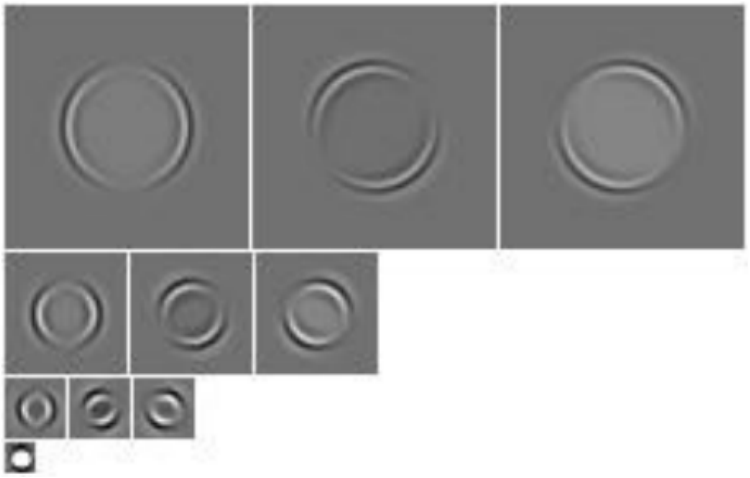


Figure 3.5: Result of steerable pyramid decomposition in three scales and three orientations. Three sub-bands images at each scale and the final lowpass image. Original from [3].

The wavelet coefficient in a sub-band is partitioned in non-overlapping blocks of size 3x3. Each block in a sub-band k, $k \in 1, 2, 3, ..., k$, is indexed by m, $m \in 1, 2, ..., Mk$. Let $\bar{\mathrm{C}}_{mkf}$ denote a vector of wavelet coefficients in the m_{th} block, the k_{th} sub-band, and the f_{th}frame. The vector $\bar{\mathrm{C}}_{mkfr}$, a vector of wavelet coefficients from the reference video, is modeled as a continuous Gaussian Scale Mixture (GSM) distribution [46], as defined in the following equation:

$$\bar{\mathrm{C}}_{mkfr} = S_{mkfr}\bar{\mathrm{U}}_{mkfr}, \tag{3.13}$$

where S_{mkfr} are independent of $\bar{\mathrm{U}}_{mkfr}$, with $\bar{\mathrm{U}}_{mkfr} \sim \mathcal{N}(0, \mathrm{K}_U kf)$. The vectors of wavelet coefficients from the distorted video are modeled in the same way, and they are defined as $\bar{\mathrm{C}}_{mkfd}$.

The SRRED is defined by the following equation:

$$SRRED_k^{M_k} = \frac{1}{FM_k}\sum_{f=1}^{F}\sum_{m=1}^{M_k} |\gamma_{mkfr}h(C'_{mkfr}|S_{mkfr} = s_{mkfr}) - \gamma_{mkfd}h(C'_{mkfd}|S_{mkfd} = s_{mkfd})|, \tag{3.14}$$

where F is the total of frames in the video, M_k is the total number of blocks in the sub-band,

$h(C'_{mkfr}|S_{mkfr} = s_{mkfr})$ and $h(C'_{mkfd}|S_{mkfd} = s_{mkfd})$ are the entropies of C'_{mkfr} and C'_{mkfd} conditioned on the maximum likelihood estimates of S_{mkfr} and S_{mkfd}. Finally, γ_{mkfr} and γ_{mkfd} are scale factors defined by the following equations:

$$\gamma_{mkfr} = \log(1 + s^2_{mkfr}) \tag{3.15}$$

$$\gamma_{mkfd} = \log(1 + s^2_{mkfd}) \tag{3.16}$$

The TRRED algorithm is based on the entropic differences of the wavelet coefficients of the current frame and the next frame, in both the reference and distorted videos. Let $\bar{\mathrm{D}}_{mkf}$ denote the vector of wavelet coefficients differences of the current frame (f) and the next frame ($f + 1$) of the m_{th} block, and the k_{th} sub-band. The block $\bar{\mathrm{D}}_{mkfr}$, from the reference video, is modeled as a continuous GSM distribution [46], as defined by the following equation:

$$\bar{\mathrm{D}}_{mkfr} = T_{mkfr}\bar{\mathrm{V}}_{mkfr}, \tag{3.17}$$

where T_{mkfr} is independent of $\bar{\mathrm{V}}_{mkfr}$, with $\bar{\mathrm{V}}_{mkfr} \sim \mathcal{N}(0, \mathrm{K}_V kf)$. The vectors of wavelet coefficients differences of the current and next frames from the distorted video are modeled in the same way, and defined as $\bar{\mathrm{D}}_{mkfd}$.

The TRRED is defined by the following equation:

$$TRRED_k^{M_k} = \frac{1}{FM_k}\sum_{f=1}^{F}\sum_{m=1}^{M_k} |\delta_{mkfr}h(D'_{mkfr}|T_{mkfr} = t_{mkfr}) - \delta_{mkfd}h(D'_{mkfd}|T_{mkfd} = t_{mkfd})|, \tag{3.18}$$

where F is the total of frames in the video, M_k is the total number of blocks in the sub-band, $h(D'_{mkfr}|T_{mkfr} = t_{mkfr})$ and $h(D'_{mkfd}|T_{mkfd} = t_{mkfd})$ are the entropies of D'_{mkfr} and D'_{mkfd}, which are conditioned on the maximum likelihood estimates of T_{mkfr} and T_{mkfd}, and δ_{mkfr} and δ_{mkfd} are scale factors defined by:

$$\gamma_{mkfr} = \log(1 + s^2_{mkfr})\log(1 + t^2_{mkfr}) \tag{3.19}$$

$$\gamma_{mkfd} = \log(1 + s^2_{mkfd})\log(1 + t^2_{mkfd}) \tag{3.20}$$

Finally the SRRED and the TRRED is combined into a single STRRED quality index. For the k_{th} sub-band, the STRRED index is defined as:

$$STRRED_k = SRRED_k \cdot TRRED_k \tag{3.21}$$

Although the video is decomposed in many sub-bands. These sub-bands have different orien-

tations and size. The STRRED is only computed for the sub-band with the vertically orientation and 1/4 of the video size, because it corresponds to the best prediction accuracy quality score performance among the different orientations and scales.

3.6 VIS3

Video Quality Assessment via Analysis of Spatial and Spatial-Temporal Slices (ViS3) is a full-reference video quality assessment method developed by Vu *et al.* [4]. This method computes video quality in two stages. First, it estimates the degradations due to spatial distortions. This stage is called ViS1. In the second stage, it estimates the degradations due to joint spatial and temporal distortions. This stage is called ViS2.

Figure 3.6 shows an overview of the ViS1 algorithm. To compute the visible distortion map, the reference frame and the distorted frame are converted to perceived luminance values using the following equation:

$$L = (\alpha + kI)^{\frac{\gamma}{3}}, \tag{3.22}$$

where I is the video frame, $\alpha = 0$, $k = 0.02874$ and $\gamma = 2.2$. These parameters are adjusted for 8-bit pixel values and sRGB displays. An error frame (ΔL) is computed using the following equation:

$$\Delta L = L_{ref} - L_{dist}, \tag{3.23}$$

where L_{ref} and L_{dist} are the reference and distorted frames converted to perceived luminance. Then, a contrast sensitivity function is applied to the reference frames and error frames using the following equation:

$$\tilde{L} = \mathscr{F}^{-1}[H(u,v) \cdot \mathscr{F}[L]], \tag{3.24}$$

where $\mathscr{F}$ and $\mathscr{F}^{-1}$ correspond to the Discrete Fourier Transform (DFT) and the Inverse DFT, and $H(u,v)$ is the Contrast Sensitivity Function in its DFT version, as defined in [41].

After this preprocessing, a local Root Mean Squared (RMS) contrast map is computed for the reference frame. The frame is divided in 16x16 blocks, with a 75% of overlap among neighboring blocks. The RMS contrast is computed for each block using the following equation:

$$C_{ref} = \frac{\tilde{\sigma}(b)}{\mu_{ref}(b)}, \tag{3.25}$$

where $\mu_{ref}(b)$ is the average pixel value of block b of $\tilde{L}$, and $\tilde{\sigma}(b)$ is the minimum of the standard deviations of the four non overlapping 8x8 blocks within block b.

Then, the local RMS contrast for the error frame (ΔL) is computed. The frame is divided in 16x16 blocks, with 75% overlap between neighboring blocks and the RMS contrast is computed

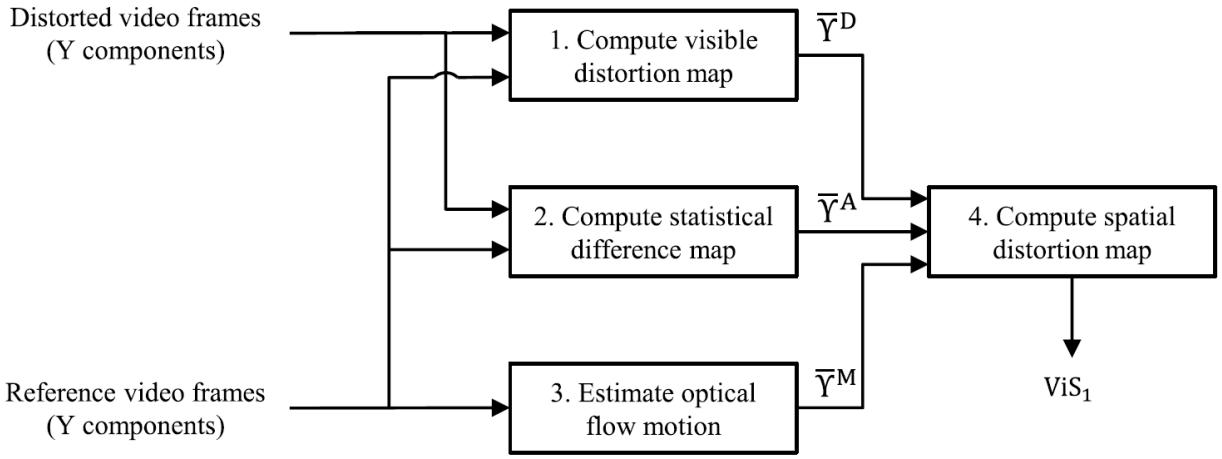


Figure 3.6: Diagram of ViS1 Algorithm. Original from [4].

for each block using the following equation:

$$C_{err}(b) = \begin{cases} \frac{\sigma_{err}(b)}{\mu_{ref}(b)} & \text{if } \mu_{ref}(b) > 0.5 \\ 0 & \text{otherwise,} \end{cases} \tag{3.26}$$

where $\mu_{ref}(b)$ is the standard deviation of the block b of the error frame (ΔL). This adapted RMS contrast algorithm takes into account the human visual system insensitivity to changes in darker areas, by making the values of darker areas equal to zero. Then, the local distortion visibility map is computed using the following equation:

$$\zeta_(b) = \begin{cases} \ln[C_{err}(b)] - \ln[C_{ref}(b)] & \text{if } \ln[C_{err}(b)] - \ln[C_{ref}(b)] > -5 \\ \ln[C_{err}(b)] + 5 & \text{if } \ln[C_{err}(b)] > -5 \geq \ln[C_{ref}(b)] \\ 0 & \text{otherwise.} \end{cases} \tag{3.27}$$

Finally, the visible distortion map is computed using a point-by-point multiplication between the local distortion visibility map (ζ) and the MSE of the reference and distorted frames, as given by the following equation:

$$\Upsilon^{D} = \zeta(b) \cdot MSE(b). \tag{3.28}$$

The statistical difference map measures the differences of the local statistics of multiscale log-gabor filter response of reference and distorted frames. The reference and distorted frames are filtered with a log-Gabor filter bank (with five scales $s \in 1, 2, 3, 4, 5$ and four orientations $o \in 1, 2, 3, 4$). The result are sets of log-Gabor subbands of the reference ($R^{s,o}$) and distorted ($\hat{R}^{s,o}$) frames. Each log-Gabor subband is divided in 16x16 blocks, with 75% overlapping between neighboring blocks, and then the standard deviation, skewness, and kurtosis are computed for each block. The statistical difference of the block b is computed as:

$$\Upsilon^A(b) = \sum_{s=1}^{5}\sum_{o=1}^{4} w_s[|\sigma^{s,o}(b) - \hat{\sigma}^{s,o}(b)| + 2|\zeta^{s,o}(b) - \hat{\zeta}^{s,o}(b)| + |\kappa^{s,o}(b) - \hat{\kappa}^{s,o}(b)|], \quad (3.29)$$

where $\sigma^{s,o}(b)$, $\zeta^{s,o}(b)$ and $\kappa^{s,o}(b)$ are the standard deviation, skewness and kurtosis of block b and subband ($R^{s,o}$), respectively. And $\hat{\sigma}^{s,o}(b)$, $\hat{\zeta}^{s,o}(b)$ and $\hat{\kappa}^{s,o}(b)$ are the standard deviation, skewness and kurtosis of block b and subband ($\hat{R}^{s,o}$).

Motion vectors are computed to model the effects of motion on video quality. In areas with larges amount of movement the distortions are less visible. Alternatively, in areas with less motion the distortions are more visible. Motion vectors are computed using the Lucas-Kanade method [47]. In this method, two matrices of motion vectors are obtained, M_v for the vertical motion and M_h for the horizontal motion. The motion magnitude matrix is computed as:

$$M = \sqrt{M_v^2 + M_h^2}. \quad (3.30)$$

The visible distortion map, statistical difference map, and the motion magnitude map are computed for a Group of Video Frames (GOF). The point-by-point average value of these maps in a GOF represents the map values for the GOF, as shown in the following equations:

$$\bar{\Upsilon}_k^D = \frac{1}{N}\sum_{\tau=1}^{N} \bar{\Upsilon}_{N(k-1)+\tau}^D \quad (3.31)$$

$$\bar{\Upsilon}_k^A = \frac{1}{N}\sum_{\tau=1}^{N} \bar{\Upsilon}_{N(k-1)+\tau}^A \quad (3.32)$$

$$\bar{\Upsilon}_k^M = \frac{1}{N}\sum_{\tau=1}^{N} \bar{M}_{N(k-1)+\tau} \quad (3.33)$$

where N is the number of frames in the GOF. ViS3 uses $N = 8$.

Finally, ViS1 is a combination of the visible distortion map, statistical difference map and motion magnitude map, as shown in the following equation:

$$ViS1 = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}\frac{[\bar{\Upsilon}_k^D(x,y)]^{\hat{\alpha}(x,y)} \cdot \bar{\Upsilon}_k^A(x,y)]^{1-\hat{\alpha}(x,y)},}{\sqrt{1+\bar{\Upsilon}_k^M(x,y)}}} \quad (3.34)$$

where $\hat{\alpha}(x,y)$ is a parameter given by the following equation:

$$\hat{\alpha}(x,y) = \frac{1}{1 + \beta_1 \cdot [\bar{\Upsilon}_k^D(x,y)]^{\beta_2}}, \tag{3.35}$$

where $\beta_1 = 0.46711$ and $\beta_2 = 0.12964$. Basically, the ViS1 score is a weighted product of the visible distortion map and the statistical difference map, divided by the motion magnitude map.

In ViS2, the frames are converted to the perceived luminance, as shown in Equation 3.22. The spatial-temporal slices are extracted from the videos, similarly to what is done in SSTS-GMSD. Let $S_x(t,y)$ and $\hat{S}_x(t,y)$ denote the vertical slices of the reference and distorted videos, and $x \in [1, W]$, where W is the video spatial width. $S_y(x,t)$ and $\hat{S}_y(x,t)$ denote the horizontal slice of the reference and distorted video, respectively, and $y \in [1, H]$, where W is the video spatial height. Figure 3.7 shows an overview of the ViS2 algorithm. First, the spatial-temporal correlation map and spatial-temporal response difference map are computed. Then, both maps are combined to compose the spatial-temporal dissimilarity map.

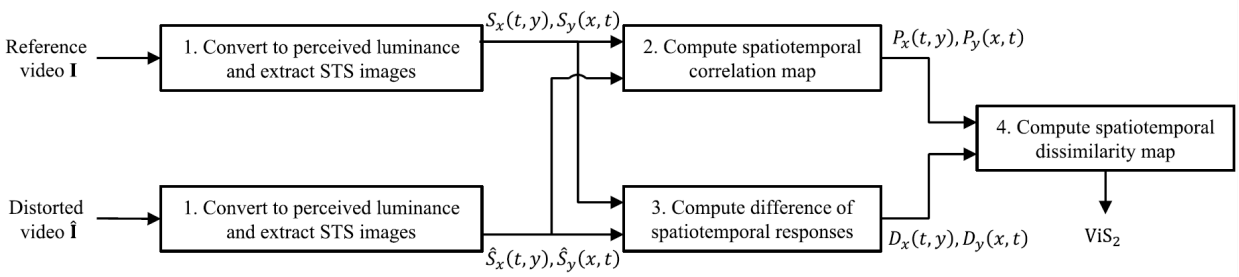


Figure 3.7: Diagram of ViS2 Algorithm. Original from [4].

To compute the spatial-temporal correlation map, the spatial-temporal slice is divided in 16x16 blocks, with a 75% overlap among neighboring blocks. Local linear correlation coefficients of the spatial-temporal slices are extracted from the reference and distorted videos, respectively, as defined in the following equation:

$$\tilde{\rho}(b) = \begin{cases} 0, & \text{if } \rho(b) < 0 \\ 1, & \text{if } \rho(b) > 0.9 \\ \rho(b) & \text{otherwise} \end{cases} \tag{3.36}$$

where ρ is the linear correlation coefficient and b is the block of the spatial-temporal slice.

A spatial-temporal response map is computed using separate 1-D filters to each dimension of the spatial-temporal slice. The spatial filter is a set of log-Gabor 1-D filters, g_s, with $s \in 1, 2, 3, 4, 5$. The frequency response of the filter is defined in the following equation:

$$G_s(\omega) = exp\left[-\frac{(ln\left|\frac{\omega}{\omega_s}\right|)^2}{2(lnB_s)^2} \right], \tag{3.37}$$

where ω_s is the center frequency of the filter g_s, B_s is the bandwidth of the filter g_s and $\omega \in [-\omega_s, \omega_s]$ is the 1-D spatial frequency.

The two temporal filters $h_z, z \in 1, 2$ are defined as:

$$h_z(t) = t^{n_z} exp(-t) \left[\frac{1}{n_z! - \frac{t^2}{(n_z+2)!}} \right], \tag{3.38}$$

where $n_1 = 6$ and $n_2 = 9$. The spatial-temporally filtered images obtained by filtering $S_x(t, y)$ are given by:

$$R_x^{s,z}(t, y) = [S_x(t, y) *^y g_s] *^t h_z, \tag{3.39}$$

The spatial-temporally filtered images obtained by filtering $S_y(t, y)$ are given by:

$$R_y^{s,z}(x, t) = [S_y(x, t) *^x g_s] *^t h_z, \tag{3.40}$$

where $*^d$ is the convolution along dimension d and $d \in x, y, t$, $s \in 1, 2, 3, 4, 5$ and $z \in 1, 2$. So, in total, there are 10 spatial-temporal filtered image for each spatial-temporal slice, $S_x(t, y)$ and $S_y(t, y)$. Similarly, $\hat{R}_x^{s,z}$ and $\hat{R}_y^{s,z}$ are computed as the spatial-temporally filtered images of $\hat{S}_x$ and $\hat{S}_y$. The absolute difference of the spatial-temporally filtered images are computed by:

$$\Delta R_x^{s,z}(t, y) = \left| R_x^{s,z}(t, y) - \hat{R}_x^{s,z}(t, y) \right|, \tag{3.41}$$

$$\Delta R_y^{s,z}(x, t) = \left| R_y^{s,z}(x, t) - \hat{R}_y^{s,z}(x, t) \right|, \tag{3.42}$$

The response difference map is computed as a natural logarithm of a weighted sum of all adjusted standard deviation maps. D_x and D_y are the log of response difference maps of the vertical and horizontal spatial-temporal slices, respectively, which are computed using the following equations:

$$D_x(t, y) = ln \left\{ 1 + A \sum_{s=1}^{5} \sum_{z=1}^{2} w_s [\tilde{\sigma}_x^{s,z}(t, y)]^2 \right\}, \tag{3.43}$$

$$D_y(x, t) = ln \left\{ 1 + A \sum_{s=1}^{5} \sum_{z=1}^{2} w_s [\tilde{\sigma}_y^{s,z}(t, y)]^2 \right\}, \tag{3.44}$$

where w_s are weights, $w_s = 0.5, 0.75, 1, 5, 6$, $A = 10^4$ is a scaling factor, $\tilde{\sigma}_x^{s,z}$ and $\sigma_y^{s,z}$ are maps of adjusted standard deviation. $\tilde{\sigma}_x^{s,z}$ and $\sigma_y^{s,z}$ are computed by:

$$\tilde{\sigma}_x^{s,z}(b) = \begin{cases} 0, & \text{if } \mu_x^{s,z} < p \\ \sigma_x^{s,z}(b)\sqrt{\frac{\mu_x^{s,z}(b)}{p+\mu_x^{s,z}(b)}} & \text{otherwise} \end{cases}, \tag{3.45}$$

$$\tilde{\sigma}_y^{s,z}(b) = \begin{cases} 0, & \text{if } \mu_y^{s,z} < p \\ \sigma_y^{s,z}(b)\sqrt{\frac{\mu_y^{s,z}(b)}{p+\mu_y^{s,z}(b)}} & \text{otherwise} \end{cases}, \tag{3.46}$$

The spatial-temporal dissimilarity value for the vertical and horizontal spatial-temporal slices are computed as the RMS value of a combination of the spatial-temporal correlation map and the log of the response difference map, as shown in the following equations:

$$\bar{\Delta}_c^{FY} = \sqrt{\frac{1}{FH}\sum_{t=1}^{F}\sum_{y=1}^{H}[D_x(t,y)\cdot\sqrt{1-P_x(t,y)}]^2}, \tag{3.47}$$

$$\bar{\Delta}_r^{XT} = \sqrt{\frac{1}{WF}\sum_{x=1}^{W}\sum_{t=1}^{F}[D_y(x,t)\cdot\sqrt{1-P_y(x,t)}]^2}, \tag{3.48}$$

where P_x and P_y are the spatial-temporal correlation maps of the vertical and horizontal spatial-temporal slices, D_x and D_y are the log of the response difference maps of the vertical and horizontal spatial-temporal slices, W is the width of the video, H is the height of the video and F is the number of frames.

The ViS2 score represents the spatial-temporal dissimilarity value and is computed with the following equation:

$$ViS2 = \sqrt{\frac{1}{W}\sum_{c=1}^{W}[\bar{\Delta_c^{FY}}]^2 + \frac{1}{H}\sum_{H}^{r=1}[\bar{\Delta_r^{XF}}]^2} \tag{3.49}$$

Finally, Figure 3.8 shows how to calculated the ViS3 scores, which is a combination of ViS1 and ViS2 scores into a single scalar. The ViS3 score represents the overall perceived video quality. And, it is a geometric mean of ViS1 and ViS2, given by:

$$ViS3 = \sqrt{ViS1 \cdot ViS2} \tag{3.50}$$

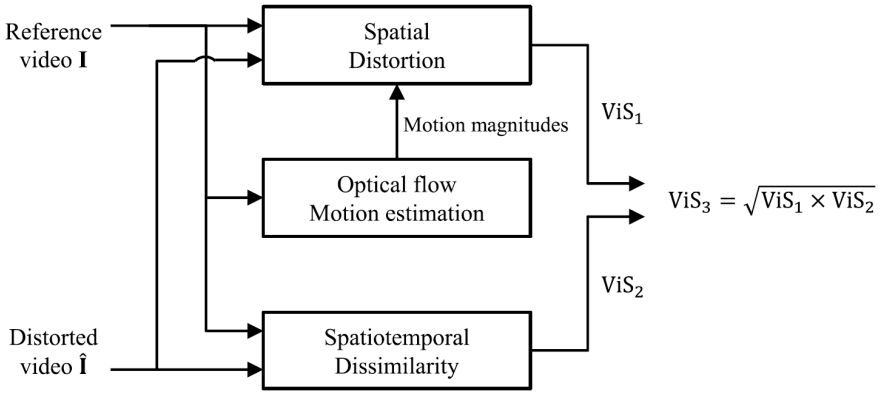


Figure 3.8: Diagram of ViS3 Algorithm. Original from [4].

4 PROPOSED FRAMEWORK AND VIDEO CLASSIFIERS

One of the challenges of video quality assessment methods is to improve the runtime performance of the algorithms [13]. The runtime performance of most of these metrics depends on the video resolution. Therefore, reducing the video spatial resolution improves their runtime. However, accuracy performance can be affected when the spatial resolution is reduced. More specifically, video distortions that affect high frequencies are the most sensitive to video spatial downsampling (MJPEG compression, packet loss and white noise). Some examples of distortions that are sensitive to downsampling include distortions introduced by MJPEG compression algorithms (e.g. blocking artifacts), packet loss and noise.

We propose a downsampling video framework that improves the runtime performance of video quality assessment methods by reducing the video resolution, but does not interfere with the accuracy performance. Before reducing the spatial resolution, the technique identifies videos that are more sensitive to this reduction. This framework is illustrated in Figure 4.1.

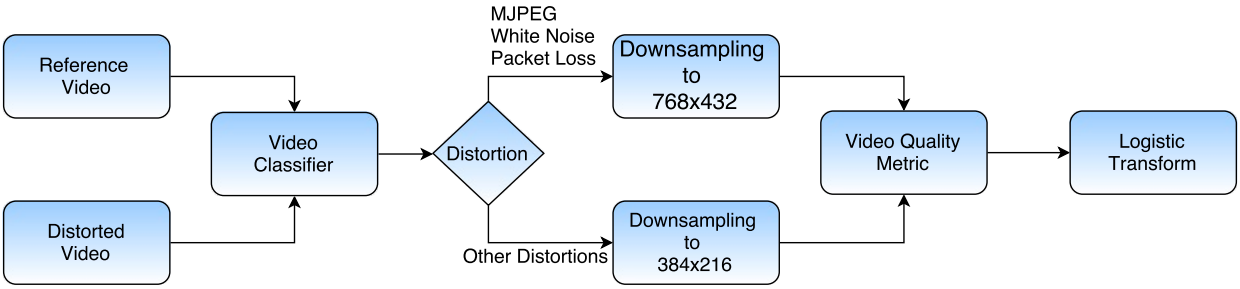


Figure 4.1: High Level Overview of the Downsampling Video Framework.

The framework has four main stages: the video classifier, the video downsampler, the quality assessment method and the logistic transform. The video classifier stage identifies the types of videos more sensitive to a spatial resolution reduction. The downsampler stage reduces the video spatial resolution, accordingly, with the goal of optimizing resolution according to the video sensitivity. Finally, the logistic transform stage adjusts the predicted score to a common scale.

We design two different video classifiers that have the goal of identifying the types of distortions in the video and choose the most adequate downsampling ratio. Video classifiers are algorithms that classify videos by analyzing one or more features. Videos can be classified according to their content, distortion type or others characteristics. Next, we describe the two video classifiers proposed in this work: a spatial activity difference classifier and a SSEQ classifier. Next, we describe the downsampling and logistic transform stage of the framework.

4.1 SPATIAL ACTIVITY DIFFERENCE CLASSIFIER

The first proposed video classifier is the spatial activity difference classifier. This classifier is based on a analysis of the Spatial Activity (SA) of the videos [48]. In our work, we have noticed that videos distorted by MJPEG compression, white noise and packet loss have a higher Spatial Activity than the videos with others types of distortions, when considering video sequences with the same content. This can be observed in Figure 4.2, where the white regions are the regions with a higher SA. Because the frames filtered with Sobel have subtle intensity changes, for a better visualization, we perform a post processing of these frames using a histogram equalization [20]. Blocking and erroneously decoded blocks leave false edges in the frame and the Sobel filter can detect them. Therefore, the idea behind this classifier is to use the SA difference between the reference and distorted videos to separate video distortions in two classes. The first class correspond to videos distorted by white noise, packet loss and MJPEG compression, while the second class corresponds to videos distorted by H.264, MPEG-2, HEVC, wavelet compressions and upscaling.

4.1.1 Spatial Activity

The SA feature is given by:

$$SA(t_n) = RMS(Sobel(I(t_n))), \tag{4.1}$$

where $I(t_n)$ is the luminance component of the video frame t_n, RMS is the RMS function over the entire image, and $Sobel$ is the Sobel Filter. The Sobel filter used in this work has two 3x3 kernels, one to detect horizontal edges, given by:

$$G_x = \begin{vmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{vmatrix}. \tag{4.2}$$

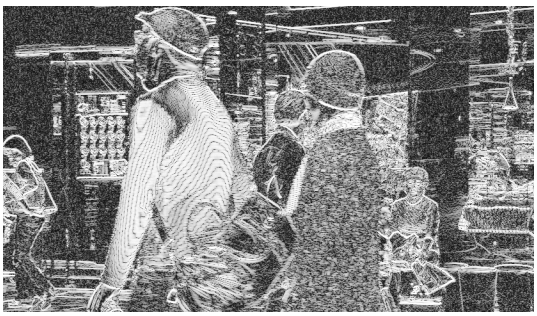
And the other kernel detects vertical edges, given by:

$$G_y = \begin{vmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{vmatrix}. \tag{4.3}$$

Then, the luminance component of the video frame is filtered with the Sobel filter and combined in the following form:

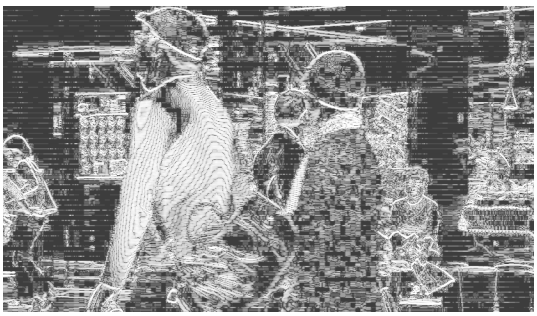
$$Sobel(t_n) = \sqrt{(G_x * I_n)^2 + (G_y * I_n)^2}, \tag{4.4}$$

(a) Reference Frame



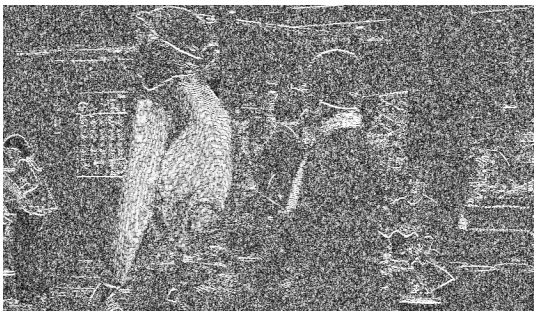
(b) Spatial Activity = 91.11

(c) MJPEG Compression



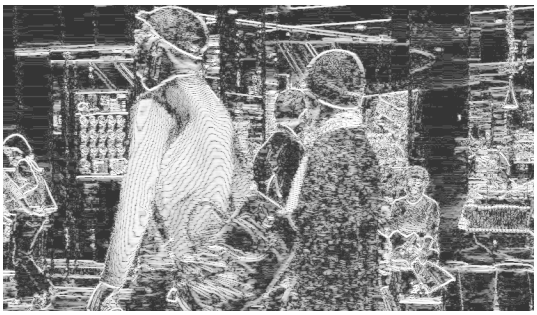
(d) Spatial Activity = 101.38

(e) White Noise



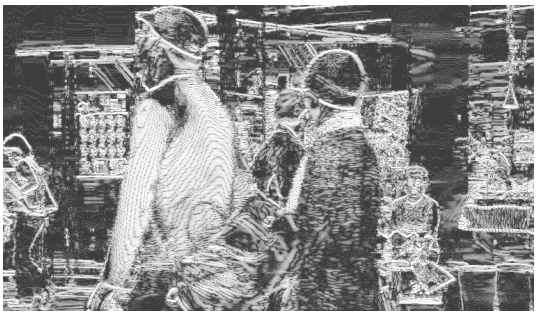
(f) Spatial Activity = 119.76

(g) H.264 Compression



(h) Spatial Activity = 88.75

(g) HEVC Compression



(h) Spatial Activity = 83.71

Figure 4.2: Frames and the result of the frame filtered with Sobel.

where $*$ is the symbol for a convolution between the kernel (G_x or G_y) and the luminance component of the video frame (I_n). We use the implementation provided by the Monitoring Of Audiovisual Quality by Key Indicators (MOAVI) group [49]. To calculate the video Spatial Activity, we average the SA values of all frames in the video.

4.1.2 Spatial Activity Difference Algorithm

Figure 4.3 shows the diagram of the proposed classification method. First, we compute the SA of the reference and distorted videos, and, then, we calculate their difference. Considering that the videos we are trying to identify have a higher (MJPEG compression and white noise) or a slightly lower space activity (packet loss) than the reference video, the SA difference must be lower than a threshold. We tested 9 values of thresholds, ranging from -4 to 4, in order to identify the best threshold. Then, we use the F1 score, which measures the accuracy of a classifier (see Appendix I), to select the best threshold. The F1 Score varies between 1 (best) and 0 (worst). Table 4.1 shows the results of using these 9 thresholds. **We tested the different thresholds using all videos of the LIVE, CSIQ, IVPL and MCL-V database.** For five different temporal resolutions, notice that a threshold lower than 1 does not provides a good performance. Results show that the best threshold value is 2. We also test the classifier performance for videos in different temporal resolutions. We noticed that it is not necessary to analyze all frames of the video sequence to obtain a good classification. It is sufficient to analyze 2 frames per second. Surprisingly, the classification results obtained by performing this temporal downsampling are better than the classification results obtained by analyzing all frames in the video.

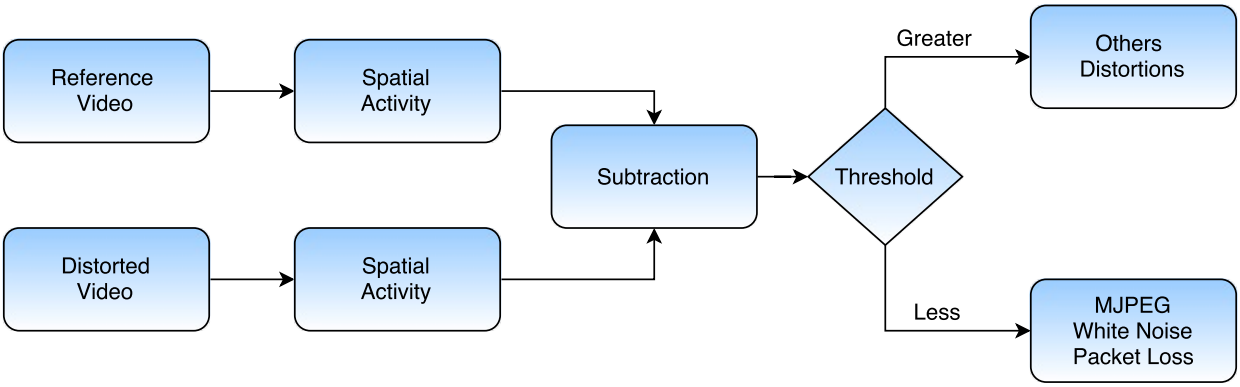


Figure 4.3: Diagram of Spatial Activity Classifier Proposed.

Table 4.1: F1 Score When Varing Threshold of Spatial Activity Difference.

FPS	Spatial Activity Difference								
	-4	-3	-2	-1	0	1	2	3	4
Original FPS	0.6145	0.6216	0.6491	0.6752	0.7032	0.7805	0.7957	0.7869	0.7368
10 FPS	0.6109	0.6216	0.6491	0.6784	0.7032	0.7830	0.7957	0.7902	0.7415
8 FPS	0.6145	0.6216	0.6491	0.6816	0.7032	0.7753	0.7997	0.7907	0.7431
4 FPS	0.6145	0.6216	0.6491	0.6752	0.7093	0.7830	0.8076	0.7858	0.7415
2 FPS	0.6145	0.6216	0.6491	0.6784	0.7062	0.7830	0.8042	0.7832	0.7411

4.2 SSEQ CLASSIFIER

The second classification method proposed in this work is based on the Spatial–Spectral Entropy-based Quality (SSEQ), which is a no-reference image quality assessment method proposed by Liu *et al.* (50). **Only the SSEQ feature extraction is used in this work.** The proposed classifier uses a Support Vector Machine (SVM) to classify the video distortions. There are four classes of video distortions: white noise, packet loss, MJPEG compression and "others distortions" (includes H.264, MPEG-2, HEVC, wavelet compressions and upscaling).

4.2.1 SSEQ Feature Extration Stage

SSEQ uses a 2-stage framework. First, the feature extraction stage extracts features from the image and uses them in a distortion classification stage. Second, a quality prediction stage estimates the video quality [51]. In this work, we only use the feature extraction stage of the SSEQ. A high-level overview of the SSEQ feature extraction stage is presented in Figure 4.4.

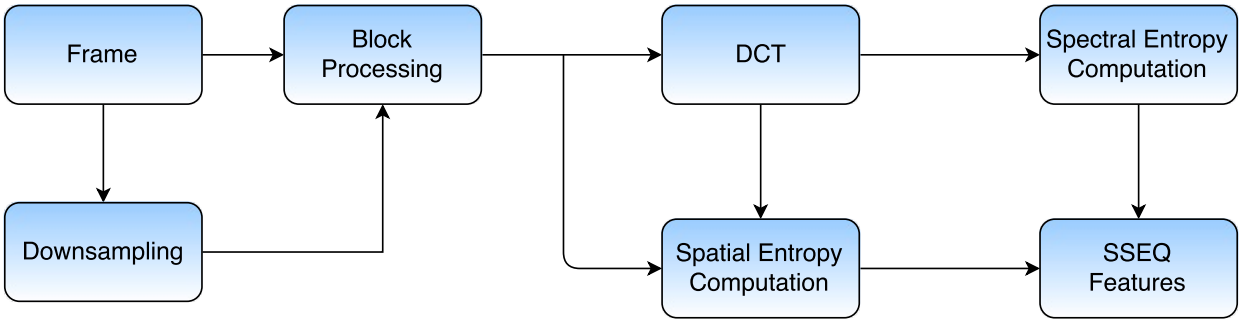


Figure 4.4: Diagram of SSEQ Feature Extraction.

The SSEQ Feature Extraction stage is described as follows. First, the frame is downsampled by a factor of 2. A bicubic interpolation method is used in the downsampling step, generating frames with 3 different sizes (1x, 0.5x and 0.25x of the original resolution). For each size, the frame is divided into 8x8 blocks, and the spatial entropy and the spectral entropy are computed for each block of the frame. The spatial entropy of a block is calculated as in the following equation:

$$E_s = -\sum_{M-1}^{x=0}\sum_{N-1}^{y=0} p(x,y) \cdot log_2(p(x,y)), \tag{4.5}$$

where MxN is the block size and $p(x,y)$ is the probability of the pixel value within the block. We compute the spatial entropy (E_s) for each block in the frame and we obtain $S(se_1, se_2, ..., se_m)$, which is a set of spatial entropy values per block, where se_i is the spatial entropy of i_{th} block. To compute the spectral entropy, we first obtain the discrete cosine transform (DCT) of each block [52]. Then, we normalize the DCT coefficients, except for the DC coefficient, as given by:

$$P(u,v) = -\frac{C(u,v)^2}{\sum_{M-1}^{x=0}\sum_{N-1}^{y=0} C(u,v)}, \tag{4.6}$$

where $C(u,v)$ is the DCT coefficient block. The spectral entropy of a block is defined by:

$$E_f = -\sum_{M-1}^{x=0}\sum_{N-1}^{y=0} P(x,y) \cdot log_2(P(x,y)), \tag{4.7}$$

We compute the spectral entropy (E_f) for each block in the frame and we obtain $Fr(fe_1, fe_2, ..., fe_m)$, which is a set of spectral entropy values per block, where fe_i is the spectral entropy of i_{th} block.

SSEQ uses a feature pooling strategy to improve the classifier performance. The pooling method used is the percentile pooling. The S and F values are sorted in ascending order and we select only the 60% central values, creating two sets: $S_c(se_{\lfloor 0.2m \rfloor}, se_{\lfloor 0.2m \rfloor+1}, ..., se_{\lfloor 0.8m \rfloor})$ and $Fr_c(fe_{\lfloor 0.2m \rfloor}, fe_{\lfloor 0.2m \rfloor+1}, ..., fe_{\lfloor 0.8m \rfloor})$. Finally the features for each scale are the mean value of S_c and of Fr_c and the skewness value of S and Fr:

$$fv = (mean(S_c), skewness(S), mean(F_c), skewness(F)). \tag{4.8}$$

The SSEQ feature vector for a frame contains 12 elements, which corresponds to the 3 frame sizes and 4 features.

4.2.2 SSEQ Classifier Algorithm

We extract the SSEQ features of each frame in the video, what generates a set of feature vectors $Fv = fv_1, fv_2, ..., fv_n$, where fv_i is the feature vector of the i_{th} frame. Then, the average of each feature will create a SSEQ feature vector for the entire video, as illustrated in Figure 4.5.

The SSEQ image quality assessment uses a SVM to classify image distortions, and the results are very good. And, SVM are already used in many video applications with good results [53, 54, 55]. So, we also use a SVM to classify video distortions.

SVM were first introduce by Cortes and Vapnik as a binary classifier [56]. Given a training set, the SVM builds an optimal separating hyperplane, defined as the one with the maximal margin of separation between the two classes. And the margin is the sum of the distances of the hyperplane from the closest point of the two classes. The multiclass problems are resolved as a combination of binary problems (57).

We had to divide the videos in groups of training and testing, to generate a SVM model. In Figure 4.6, we shown an overview of the SSEQ classifier algorithm. First, we divide the videos in a training set. Then, we extract the SSEQ features of all training videos and label them according to their distortion. Then, we train the SVM with these parameters (SSEQ features and video

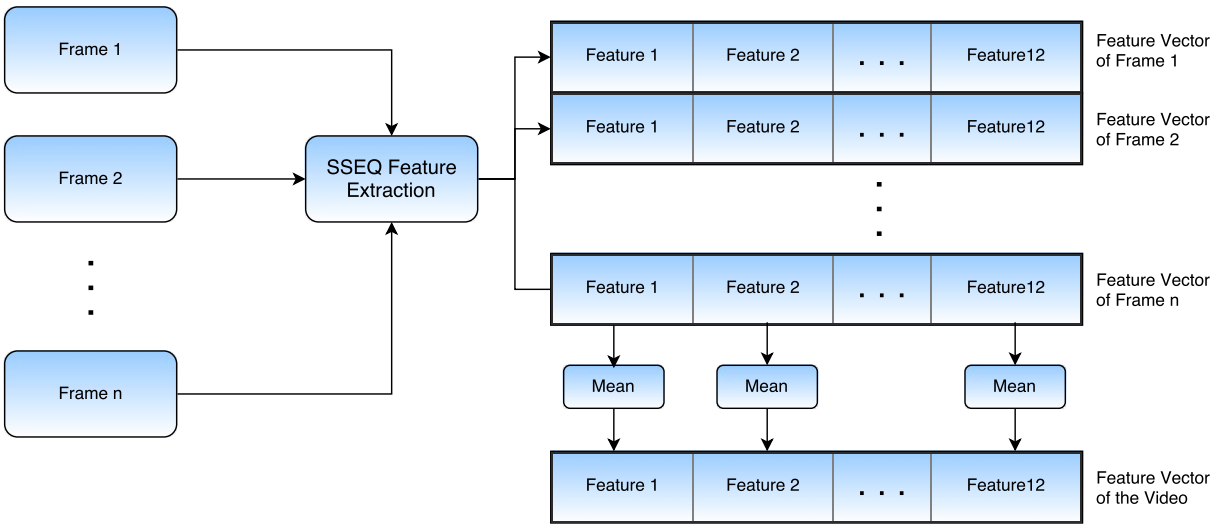


Figure 4.5: SSEQ Video Feature Extraction.

Table 4.2: F1 Score of SSEQ classifier when reducing FPS.

	FPS				
	Original	10	8	4	2
F1 Score	0.7999	0.8021	0.8016	0.8043	0.7981

distortion labels). The SVM creates a model that is used to classify others videos in the test set. In other words, SSEQ features are extracted from a test video and used as parameters of the trained SVM, what gives the estimated video distortion labels.

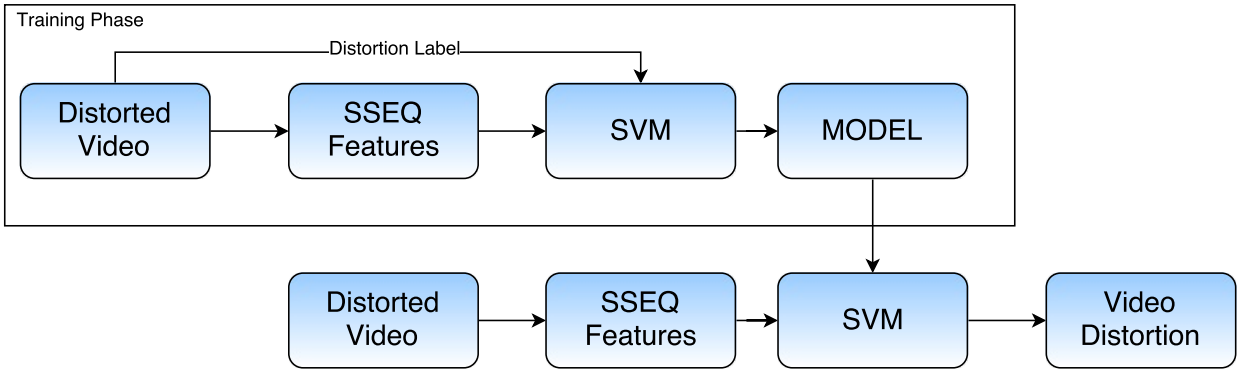


Figure 4.6: High level overview of SSEQ algorithm for traning and test.

We tested the SSEQ classifier using all videos from LIVE, CSIQ, IVPL and MCL-V databases. We divided the videos in training and testing sets. The training set consists of 80% of all videos, and the testing set consist of 20% of all videos. There are no content overlapping in the sets. Also to analyze the computational time of the SSEQ features extraction stage, we test the SSEQ video classifier with different frame rates, as shown in Table 4.2. We use the F1-Score to measure the accuracy of the classifier. We notice that even if we reduce the video temporal resolution to 2 FPS, the performance of SSEQ classifier is practically unaltered.

Table 4.3 shows the mean confusion matrix for 1000 simulations of the SSEQ classifier com-

puted with videos with 2 FPS. The confusion matrix shows the proportion of positive and negative results of the video classifier (see Appendix I). These results show that the classifier can correctly identify videos distorted by MJPEG compression and white noise. But, it struggles to identify videos distorted by packet loss. We believe, packet loss is a complicated distortion to identify, because it is a localized distortion that does not affect all frames in the same way. Table 4.4 shows the median of precision, recall, and F1 score over 1000 simulations of the SSEQ classifier. In more than half of these simulations the SSEQ classifier identifies all videos distorted by white noise and MJPEG compression, but only 60% of the videos distorted by packet loss. Also, from all videos classified as distorted by packet loss, only 37% are actually videos distorted by packet loss. Although this classifier cannot identify packet loss, we will shown in next sections that the performance of this classifier is satisfactory for our propose.

Table 4.3: Mean Confusion Matrix of SSEQ Distortion Classifier through 1000 simulations.

		Label			
		Others	MJPEG	White Noise	Packet Loss
Estimated	Others	0.8891	0.0149	0.000	0.0960
	MJPEG	0.0622	0.9378	0.0000	0.0000
	WhiteNoise	0.0000	0.0000	0.9935	0.0065
	Packet Loss	0.6031	0.0005	0.0050	0.3914

Table 4.4: Evaluation of SSEQ Distortion Classifier. Median values of Precision, Recall and F1 Score through 1000 simulations.

	Precision	Recall	F1 Score
Others	0.9000	0.7901	0.799
MJPEG	1.000	1.000	
White Noise	1.000	1.000	
Packet Loss	0.3750	0.6000	

4.3 ADJUSTED PREDICT SCORE

One of our goals is to test the performance of video quality assessment methods spatial resolution of the videos is reduced. In this work, we noticed that all metrics are sensitive to video resolution reduction. When the spatial resolution is reduced, the video quality is lower than for a video in the original resolution. For example, in Figure 4.7, we notice that for videos in 384x216 resolution, ViS3's predicted scores vary from 10^{-2} to 10, but, for the same videos in 768x432 resolution, they vary from 10^{-1} to 10. So, when we evaluate quality using a reduced version of a video, the quality score tends to be lower than the score obtained with the original resolution.

In this work, we reduce the videos to 768x432 or 384x216, depending on the video distortion. **The 768x432 resolution were chosen, because it is a resolution commonly used and it is the resolution for LIVE Database. The LIVE database has the lower spatial resolution between the databases used in this work.** In order for the predicted scores have the same pre-

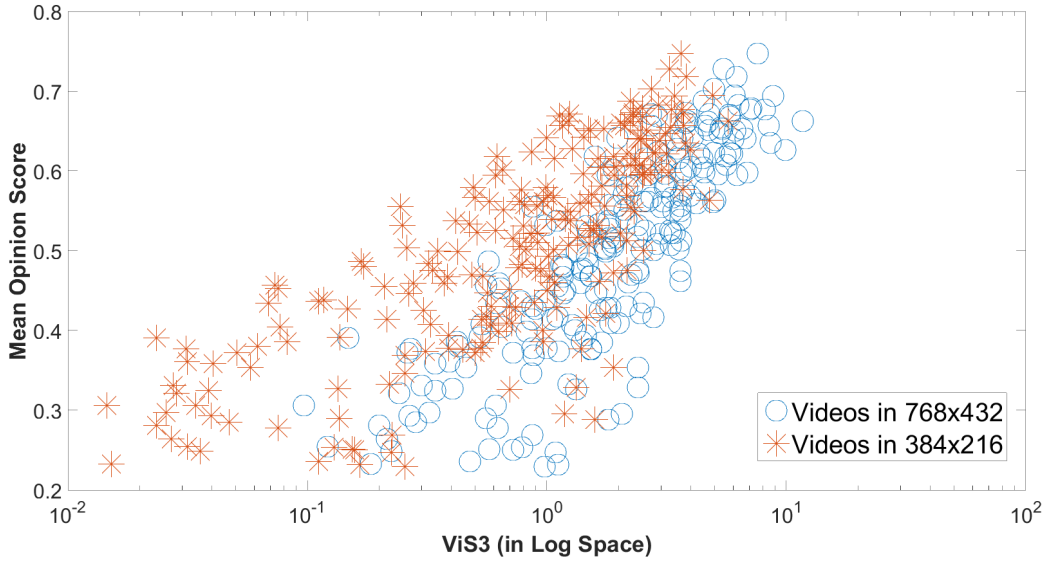


Figure 4.7: Plot of ViS3 *versus* Mean Opnion Score. The videos used were from CSIQ Database and they were downsampling to 768x432 and 384x216.

cision, independent of the video resolution, we use a logistic function to normalize the predicted scores. Figure 4.8 shows an overview operations performed to adjust the predicted score. As mentioned earlier, the downsampling operation is performed using the bicubic interpolation in each frame [58]. This technique generally gives better results than other interpolation algorithms, like nearest neighbor and bilinear. Also, bicubic interpolation is the standard used in softwares like ffmpeg, Matlab and Adobe Photoshop [20].

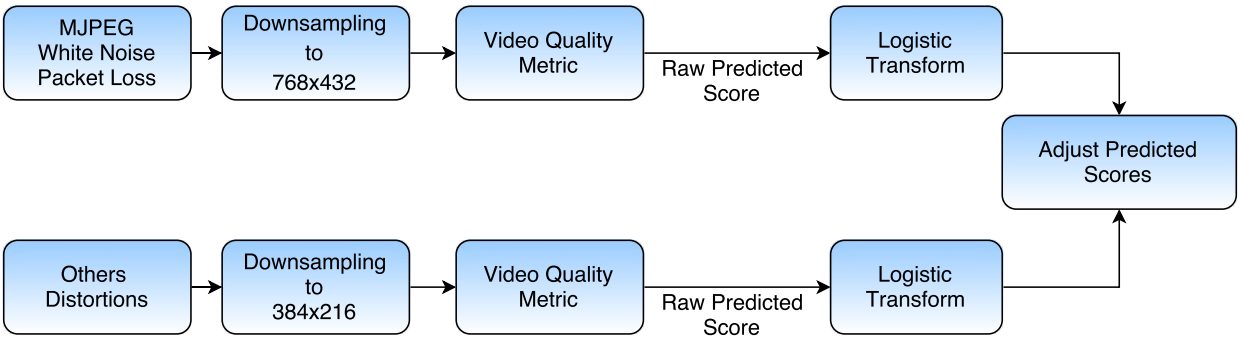


Figure 4.8: High Level Overview of the Adjustment of Predicted Score.

We chose the logistic function to normalize the scores from different resolutions because it is the scaling transform recommended by the Video Quality Experts Group (VQEG) to remove non-linearity behaviors and facilitate the comparison among scores provided by different video quality assessment methods [59]. The logistic function is given by the following equation:

$$f(x) = \frac{\tau_1}{1 + \tau_2 \cdot exp(-\tau_3 \cdot x)}, \tag{4.9}$$

where x is the predicted score and τ_1, τ_2 and τ_3 are parameters used to provide the best fit of

the predicted score to the subjective score. After using the logistic function, each score is in the same scale as the subjective scores, therefore, we can compare the scores for videos in different resolutions.

5 RESULTS

As mentioned in previous chapters, one of the challenges for video quality assessment methods is the runtime performance [12, 13]. These methods are becoming more complex over the decades, with some of the algorithms being very slow to compute. Reducing the spatial resolution of a video is one of the simplest and fastest methods to improve the running time of a video quality assessment method. However, reducing the video resolution can introduce new distortions and affect the quality prediction accuracy. In this work, we analyze the effects of reducing the video resolution on video quality assessment methods.

Based on the idea of reducing the spatial resolution, we proposed a framework, which improves the running time performance of video quality assessment methods, without affecting its accuracy performance. In the previous chapters, we detailed this framework and proposed two video classifiers that can be used in the framework. In this chapter, we study which distortions are more sensitive to the video resolution reduction. We use the proposed classifiers to identify these videos, perform a video resolution reduction and tested the framework using both classifiers. Also, we compute the running time of the video quality assessment methods to compare it to the running times of the same method using the proposed framework.

This chapter is divided in two sections. In the first section, we describe the experimental setup used in our experiments. In the second section, we discuss our experimental results.

5.1 EXPERIMENTAL SETUP

The experiments were performed on an Intel i7-4790 processor at 3.60GHz. For all video quality assessment methods, except for the SSTS-GMSD, the Matlab code was provided by the authors. To evaluate the performance of each algorithm, we chose four video quality databases, which have different resolutions and distortions.

To evaluate the accuracy performance of the video quality assessment methods we use four video quality databases, previously described in Section 2. Table 5.1 shows the specifications of the four video databases. Notice that the databases are very different form each other. IVP and MCL-V have videos in higher resolution than LIVE and CSIQ. However, CSIQ has more distortions and different temporal resolution. In the real world, videos do not have the same resolution or same distortions, so it is important these differences in the databases to better represent the videos in the real world.

Table 5.1: Comparison Between Video Databases

	Reference Videos	Total Videos	Spatial Resolution	Temporal Resolution	Type of Distortions
LIVE	10	150	768x432	25, 50	MPEG-2, H.264 Packet Loss
CSIQ	12	216	832x480	24, 25, 30 50, 60	H.264, H.265, Wavelet, MJPEG, White Noise, Packet Loss
IVP	10	128	1920x1088	25	MPEG-2, H.264, Wavelet, Packet Loss
MCL-V	12	96	1920x1080	24, 25, 30	H.264, Upscaling

5.2 EXPERIMENTAL RESULTS

In this section, we present all the tests we perform in this work. First, we study the effects on video quality assessment methods when reducing the video spatial resolution. Then, we study how different video distortions affect the video quality assessment methods performance, when reducing the video spatial resolution. From that results, we obtain the necessary information to proposed our framework, and we present the results of our framework with two video classifier. Finally, we present additional tests on the performance of video quality assessment methods when reducing the temporal resolution.

To study how the video resolution can affect the method accuracy performance, we perform some tests with videos of four different databases using different video quality assessment methods with different video spatial resolutions. To evaluate the prediction quality accuracy of each video quality assessment method, we calculate the SCC between subjective scores provided by the databases and objectives scores predicted by these methods [59]. The SCC values vary from 1 (best) to 0 (worst). When the SCC is closer to 1, it means that the predicted scores given by the quality assessment methods are similar to the subjective scores provided by the database (see Appendix II).

It is worth mentioning again that, for some video quality assessment methods, reducing the video resolution can cause a decrease in prediction quality accuracy. Tables 5.2- 5.5 shows the SCC values for a set of video quality assessment methods, tested on the LIVE, CSIQ, IVP, and MCL-V databases. Notice that for the IVP and MCL-V databases, there is an improvement on accuracy performance when the video resolution reduces, but for LIVE and CSIQ Database the accuracy performance decreases. The accuracy performance of PSNR, SSIM and GMSD increases for the LIVE database, but only the accuracy performance of SSIM increases for the CSIQ Database. It is worth pointing out that the CSIQ Database has more distortions than the others databases, while MCL-V and IVP databases have videos with a higher resolution (1920x1080 for MCL-V and 1920x1088 for IVP).

In summary, the accuracy performance of the methods vary according to the databases, i.e. according to the content and type of distortions in each database. So, to verify if the video

Table 5.2: Spearman Correlation Coefficient on the LIVE database.

	768x432	640x360	480x270	384x216
ViS3	**0.8168**	0.7995	0.7786	0.7595
STRRED	**0.8007**	0.7902	0.7691	0.7570
SSTS-GMSD	**0.8389**	0.8255	0.7988	0.7777
GMSD	0.7262	0.7354	0.7348	0.7264
SSIM	0.5251	0.6131	0.6696	**0.7002**
PSNR	0.5233	0.5750	0.6084	**0.6285**

Table 5.3: Spearman Correlation Coefficient on the CSIQ database.

	768x432	640x360	480x270	384x216
ViS3	**0.8581**	0.8574	0.8271	0.7897
STRRED	**0.8035**	0.7881	0.7666	0.7424
SSTS-GMSD	**0.8457**	0.8411	0.7950	0.7458
GMSD	**0.8449**	**0.8449**	0.8107	0.7729
SSIM	0.6236	0.6566	0.7016	**0.7241**
PSNR	**0.5896**	0.5853	0.5696	0.5585

Table 5.4: Spearman Correlation Coefficient on the IVP database.

	1920x1080	1280x720	960x540	768x432	640x360	480x270	384x216
ViS3	0.8023	0.8822	0.8995	0.8968	**0.9094**	0.8981	0.8912
STRRED	0.7378	**0.7426**	0.7308	0.7296	0.7177	0.7160	0.7144
SSTS-GMSD	0.7560	0.8236	0.8683	0.8830	**0.8880**	**0.8880**	0.8722
GMSD	0.6924	0.7973	0.8493	0.8671	0.8791	**0.8833**	0.8672
SSIM	0.3739	0.4749	0.5277	0.5677	0.6077	0.6629	**0.6965**
PSNR	0.6566	0.7050	0.7312	0.7707	0.8086	0.8333	**0.8374**

Table 5.5: Spearman Correlation Coefficient on the MCL-V database.

	1920x1080	1280x720	960x540	768x432	640x360	480x270	384x216
ViS3	0.6361	0.6902	0.7167	0.7307	0.7408	0.7516	**0.7674**
STRRED	0.7433	0.7433	0.7738	0.7938	0.7964	0.8035	**0.8177**
SSTS-GMSD	0.6855	0.7395	0.7575	0.7768	0.7850	0.7989	**0.8000**
GMSD	0.6449	0.7068	0.7234	0.7369	0.7483	0.7625	**0.7708**
SSIM	0.4018	0.5408	0.6062	0.6415	0.6632	0.6961	**0.7105**
PSNR	0.4640	0.5328	0.5791	0.6097	0.6347	0.6668	**0.6876**

distortion has influence on the accuracy performance of the methods, we analyze the data in groups separated by the types of distortion. These results are presented in Tables 5.6- 5.9.

Table 5.6 shows the accuracy performance of each method when the video resolution is reduced, for the CSIQ Database. the accuracy performance of PSNR and SSIM improves when the video resolution is reduced. For PSNR, the improvement is 25.47% for videos distorted by H.264 compression, 33.30% for videos distorted by MPEG2 compression and 12.49% for videos distorted by packet loss. For SSIM, the improvement is 4.23% for videos distorted by H.264 compression, 21.37% for videos distorted by MPEG2 compression and 32.88% for videos distorted

Table 5.6: Spearman Correlation Coefficient on the LIVE database.

Metric	Distortion	Video Resolution			
		768x432	640x360	480x270	384x216
ViS3	H.264	**0.7685**	0.7598	0.7371	0.7405
	MPEG2	0.7362	0.7639	**0.7806**	0.7774
	Packet Loss	**0.8372**	0.8298	0.8195	0.8270
STRRED	H.264	0.8193	0.8223	0.8113	**0.8304**
	MPEG2	**0.7193**	0.6986	0.6703	0.6769
	Packet Loss	**0.7934**	0.7904	0.7693	0.7510
SSTS-GMSD	H.264	**0.7974**	0.7788	0.7615	0.7623
	MPEG2	0.8125	0.8121	0.8362	**0.8470**
	Packet Loss	**0.8151**	0.8148	0.8044	0.7963
GMSD	H.264	**0.6471**	0.6452	0.6302	0.6349
	MPEG2	**0.6915**	0.6836	0.6725	0.6778
	Packet Loss	0.7457	0.7626	0.7775	**0.7800**
SSIM	H.264	0.6561	**0.7171**	0.6961	0.6839
	MPEG2	0.5609	0.6112	0.6679	**0.6808**
	Packet Loss	0.5151	0.5716	0.6337	**0.6845**
PSNR	H.264	0.4730	0.5424	0.5856	**0.5934**
	MPEG2	0.3830	0.4510	0.4825	**0.5106**
	Packet Loss	0.5799	0.6294	**0.6550**	0.6523

by packet loss. For GMSD, the accuracy performance improves for videos distorted by packet loss (4.59% of improvement), but decreases for videos distorted by H.264 and MPEG2 compression by 1.88% and 2.14%, respectively. For SSTS-GMSD, the accuracy performance improves for videos distorted by MPEG2 compression, with an improvement of 4.25%, but decreases for videos distorted by H.264 compression and packet loss by 4.40% and 2.30% respectively. For STRRED, the accuracy performance improves for videos distorted by H.264 Compression by 1.35%, but decreases for videos distorted by MPEG2 Compression and packet loss by 5.89% and 5.35% respectively. Finally, for ViS3, the accuracy performance improves for videos with MPEG2 Compression by 5.61%, but decreases for videos distorted by H.264 Compression and Packet Loss by 3.64% and 1.22% respectively.

Table 5.7 shows the accuracy performance of each method when the video resolution is reduced, for the CSIQ Database. Notice that the accuracy performance of ViS3, SSTS-GMSD, SSIM and PSNR increases when the resolution is reduced, for videos distorted by H.264, H.265 and Wavelet Compression. The gain for videos distorted by H.264 Compression is 0.58% for ViS3, 1.42% for SSTS-GMSD, 6.74% for SSIM and 7.08% for PSNR. The gain for videos distorted by H.265 Compression is 1.24% for ViS3, 0.06% for SSTS-GMSD, 9.13% for SSIM and 5.60% for PSNR. The gain for videos distorted by Wavelet Compression is 2.89% for ViS3, 1.88% for SSTS-GMSD, 5.30% for SSIM and 7.78% for PSNR. The accuracy performance of all methods decreases for videos distorted by MJPEG compression, packet loss and white noise. For all methods, the worst accuracy performance corresponds to the resolution 384x216 for videos distorted by MJPEG compression. The decrease in accuracy performance for videos distorted

Table 5.7: Spearman Correlation Coefficient on the CSIQ database.

Metric	Distortion	Video Resolution			
		768x432	640x360	480x270	384x216
ViS3	H.264	0.9300	0.9400	**0.9421**	0.9354
	H.265	0.9331	0.9413	**0.9537**	0.9447
	MJPEG	0.7776	**0.7817**	0.6829	0.5748
	Packet Loss	**0.8381**	0.8319	0.8051	0.7907
	Wavelet	0.9161	0.9238	0.9398	**0.9426**
	White Noise	**0.9210**	0.9192	0.9081	0.8929
STRRED	H.264	**0.9768**	0.9753	0.9686	0.9637
	H.265	0.9112	**0.9238**	0.9230	0.9202
	MJPEG	**0.5964**	0.4839	0.1611	0.0625
	Packet Loss	0.8468	**0.8551**	0.8435	0.8353
	Wavelet	**0.9457**	0.9351	0.9290	0.9266
	White Noise	**0.9192**	0.9079	0.8772	0.8528
SSTS-GMSD	H.264	0.9277	0.9292	**0.9318**	0.9282
	H.265	0.9454	**0.9552**	0.9480	0.9472
	MJPEG	**0.8893**	0.8885	0.8528	0.7840
	Packet Loss	**0.8108**	0.8031	0.7743	0.7495
	Wavelet	0.8705	0.8718	**0.8754**	**0.8754**
	White Noise	**0.8847**	0.8708	0.8072	0.7529
GMSD	H.264	0.9441	0.9503	**0.9532**	0.9503
	H.265	0.9408	**0.9439**	0.9326	0.9279
	MJPEG	0.8994	**0.9192**	0.8741	0.7586
	Packet Loss	**0.8631**	0.8577	0.8512	0.8111
	Wavelet	**0.8764**	0.8762	0.8623	0.8610
	White Noise	**0.9094**	0.9068	0.8808	0.8391
SSIM	H.264	0.8903	0.9143	0.9485	**0.9503**
	H.265	0.8654	0.9053	0.9290	**0.9444**
	MJPEG	**0.8373**	0.8456	0.8041	0.7797
	Packet Loss	**0.8556**	0.8535	0.8517	0.8492
	Wavelet	0.8154	0.8317	**0.8597**	0.8587
	White Noise	0.9269	**0.9292**	0.9223	0.9122
PSNR	H.264	0.8687	0.8914	0.9218	**0.9302**
	H.265	0.8556	0.8782	0.8986	**0.9035**
	MJPEG	**0.4672**	0.4296	0.2847	0.1694
	Packet Loss	**0.8481**	0.8409	0.8322	0.8245
	Wavelet	0.7941	0.8265	0.8479	**0.8559**
	White Noise	0.9048	0.9060	**0.9079**	0.8947

by MJPEG compression is 26.08% for ViS3, 89.51% for STRRED, 9.46% for SSTS-GMSD, 16.54% for GMSD, 6.89% for SSIM and 63.75% for PSNR.

Table 5.8 shows the accuracy performance for the IVP Database. Notice that the accuracy performance of videos distorted by Packet Loss decreases for the video methods (ViS3, STRRED and SSTS-GMSD) for most spatial resolutions. Nevertheless, it improves when the video resolution is 768x432. For GMSD, the best accuracy performance is obtained for videos distorted by Packet Loss when the video resolution is 640x360. The accuracy performance for videos dis-

Table 5.8: Spearman Correlation Coefficient on the IVP database.

Metric	Distortion	Video Resolution						
		1920x1080	1280x720	960x540	768x432	640x360	480x270	384x216
ViS3	Wavelet	0.9186	0.9212	0.9132	0.9253	**0.9368**	0.9212	0.9306
	H.264	0.8477	0.8735	0.8681	0.8538	0.8787	0.8864	**0.8906**
	MPEG2	0.7918	0.8274	0.8527	0.8509	0.8394	0.8558	**0.8658**
	Packet Loss	0.7504	**0.8057**	0.7947	0.7750	0.7526	0.7280	0.6962
STRRED	Wavelet	0.8554	0.8839	0.8932	**0.8914**	0.8910	0.8870	0.8843
	H.264	**0.8614**	0.8505	0.8493	0.8477	0.8435	0.8454	0.8458
	MPEG2	**0.6752**	0.6400	0.6427	0.6400	0.6360	0.6489	0.6570
	Packet Loss	0.6650	**0.7028**	0.6765	0.6661	0.6650	0.6158	0.5977
SSTS-GMSD	Wavelet	0.8216	0.8509	0.8710	0.8763	0.8812	0.8968	**0.9075**
	H.264	0.8463	0.8587	0.8651	0.8726	0.8809	0.8841	**0.8876**
	MPEG2	0.7824	0.8073	0.8318	0.8443	0.8501	0.8714	**0.8812**
	Packet Loss	0.7663	0.7871	**0.7937**	0.7750	0.7603	0.7154	0.6907
GMSD	Wavelet	0.8229	0.8763	0.8941	0.8790	0.9008	**0.9110**	0.9101
	H.264	0.8630	0.8780	0.8826	0.8856	0.8880	0.8897	**0.8902**
	MPEG2	0.8283	0.8256	0.8220	0.8211	**0.8291**	0.8251	0.8283
	Packet Loss	0.7121	0.8150	0.8413	0.8407	**0.8467**	0.8128	0.7849
SSIM	Wavelet	0.7846	0.8029	0.8060	0.8140	0.8287	**0.8492**	0.8407
	H.264	0.6636	0.7752	0.8146	0.8310	0.8407	**0.8462**	0.8418
	MPEG2	0.5884	0.6436	0.6574	0.6783	**0.6836**	0.6690	0.6529
	Packet Loss	0.0482	0.1910	0.2638	0.3136	0.3985	0.4729	**0.5161**
PSNR	Wavelet	0.8598	0.8603	0.8594	0.8759	0.8723	**0.8914**	0.8905
	H.264	0.8218	0.8488	0.8681	0.8765	0.8811	**0.8901**	**0.8901**
	MPEG2	0.6948	0.7130	0.7384	0.7620	0.7539	0.7673	**0.7749**
	Packet Loss	0.6393	0.6793	0.7115	0.7340	0.7630	0.7822	**0.8051**

Table 5.9: Spearman Correlation Coefficient on the MCL-V database.

Metric	Distortion	Video Resolution						
		1920x1080	1280x720	960x540	768x432	640x360	480x270	384x216
ViS3	H.264	0.5868	0.6625	0.6960	0.7273	0.7445	0.7655	**0.7780**
	Upscaling	0.6890	0.7043	0.7245	0.7283	0.7304	0.7333	**0.7440**
STRRED	H.264	0.7716	0.7716	0.7807	0.8055	0.8010	0.8035	**0.8139**
	Upscaling	0.7040	0.7040	0.7436	0.7732	0.7714	0.7858	**0.8123**
SSTS-GMSD	H.264	0.6921	0.7189	0.7423	0.7769	0.7867	0.7950	**0.8058**
	Upscaling	0.6806	0.7477	0.7629	0.7785	0.7778	0.7843	**0.7960**
GMSD	H.264	0.6452	0.7057	0.7332	0.7484	0.7635	0.7685	**0.7905**
	Upscaling	0.6376	0.6789	0.6898	0.7021	0.6988	0.7100	**0.7292**
SSIM	H.264	0.3545	0.5375	0.6172	0.6533	0.6764	0.7222	**0.7360**
	Upscaling	0.4400	0.5300	0.5931	0.6242	0.6408	0.6616	**0.6754**
PSNR	H.264	0.4215	0.5115	0.5634	0.5872	0.6203	0.6651	**0.6761**
	Upscaling	0.4925	0.5382	0.5851	0.6216	0.6402	0.6537	**0.6834**

torted by Wavelet and H.264 Compression improves for almost all methods, except for STRRED for videos distorted by H.264.

Table 5.9 shows the accuracy performance for each method when the video resolution is re-

Table 5.10: Spearman Correlation Coefficient of VIS3 on all databases.

Distortion	Video Resolution			
	768x432	640x360	480x270	384x216
Wavelet	0.8864	0.8988	0.9093	**0.9125**
H.264	0.8354	0.8469	**0.8479**	0.8473
H.265	0.9331	0.9413	**0.9537**	0.9447
MJPEG	0.7776	**0.7817**	0.6829	0.5748
MPEG2	0.8561	0.8766	**0.8851**	0.8848
Upscaling	0.7283	0.7304	0.7333	**0.7440**
Packet Loss	**0.8174**	0.8179	0.8087	0.8028
White Noise	**0.9210**	0.9192	0.9081	0.8929

Table 5.11: Spearman Correlation Coefficient of STRRED on all databases.

Distortion	Video Resolution			
	768x432	640x360	480x270	384x216
Wavelet	0.8505	0.8499	0.8548	**0.8638**
H.264	0.8822	0.8816	0.8810	**0.8859**
H.265	0.9112	**0.9238**	0.9230	0.9202
MJPEG	**0.5964**	0.4839	0.1611	0.0625
MPEG2	**0.7993**	0.7891	0.7874	0.7935
Upscaling	0.7732	0.7714	0.7858	**0.8123**
Packet Loss	0.7881	**0.7906**	0.7793	0.7672
White Noise	**0.9192**	0.9079	0.8772	0.8528

duced for the MCL-V Database. In this database, the accuracy performance of all methods are improved when the video resolution is reduced. For videos distorted by H.264 Compression the improvement is 32.58% for ViS3, 5.48% for STRRED, 16.43% for SSTS-GMSD, 22.52% for GMSD, 107.62% for SSIM and 60.40% for PSNR. And for videos distorted by Upscaling the improvement is 7.98% for ViS3, 15.38% for STRRED, 16.96% for SSTS-GMSD, 14.36% for GMSD, 53.50% for SSIM and 38.76% for PSNR.

From these results, we can notice that for all methods, the accuracy performance decreases when videos distorted by MJPEG compression and white noise have their resolution reduced. For videos distorted by packet loss, the accuracy performance of most methods is better when the video resolution is 640x360 or above.

To better evaluate the prediction quality accuracy of each method for each type of video distortions, we combine all videos with the same distortion type from all databases into the same group. The accuracy performance of each method is evaluated by comparing the predicted scores with the subjective scores provided by the databases. Unfortunately, each database provides subjective scores using different experimental methodologies, different subject pools, and different physical environments. As described in the previously subsection, the LIVE and IVP databases provide a DMOS for each distorted video, while the CSIQ and MCL-V databases provide a MOS of each distorted video. So, to properly combine these subjective scores into a single scale, we use the Iterated Nested Least-Squares Algorithm (INLSA) [60]. We make the video resolution

Table 5.12: Spearman Correlation Coefficient of SSTS-GMSD on all databases.

Distortion	Video Resolution			
	768x432	640x360	480x270	384x216
Wavelet	0.8035	0.8109	0.8234	**0.8367**
H.264	0.8520	0.8520	0.8539	**0.8583**
H.265	0.9454	**0.9552**	0.9480	0.9472
MJPEG	**0.8893**	0.8885	0.8528	0.7840
MPEG2	0.8776	0.8905	0.9032	**0.9074**
Upscaling	0.7785	0.7778	0.7843	**0.7960**
Packet Loss	0.7785	**0.7876**	0.7794	0.7703
White Noise	**0.8847**	0.8708	0.8072	0.7529

Table 5.13: Spearman Correlation Coefficient of GMSD on all databases.

Distortion	Video Resolution			
	768x432	640x360	480x270	384x216
Wavelet	0.8430	0.8618	0.8655	**0.8698**
H.264	0.8405	0.8439	0.8482	**0.8516**
H.265	0.9408	**0.9439**	0.9326	0.9279
MJPEG	0.8994	**0.9181**	0.8741	0.7506
MPEG2	0.8282	0.8326	0.8360	**0.8425**
Upscaling	0.7031	0.7057	0.7207	**0.7289**
Packet Loss	0.7828	0.7941	**0.7979**	0.7949
White Noise	**0.9094**	0.9068	0.8808	0.8347

Table 5.14: Spearman Correlation Coefficient of SSIM on all databases.

Distortion	Video Resolution			
	768x432	640x360	480x270	384x216
Wavelet	0.7981	0.8197	0.8469	**0.8497**
H.264	0.7962	0.8170	0.8344	**0.8398**
H.265	0.8654	0.9053	0.9290	**0.9444**
MJPEG	0.8373	**0.8456**	0.8041	0.7797
MPEG2	0.7682	0.7954	0.8160	**0.8190**
Upscaling	0.6242	0.6408	0.6616	**0.6754**
Packet Loss	0.5446	0.5732	0.6112	**0.6517**
White Noise	0.9269	**0.9292**	0.9223	0.9122

vary from 768x432 to 384x216. The lowest video resolution among all databases is 768x432.

Tables 5.10-5.15 show the results of each method's accuracy performance separated by video distortion. Notice that for the video methods, ViS3, STRRED and SSTS-GMSD, reducing the video resolution decreases their accuracy performance, for videos distorted by MJPEG, packet loss and white noise.

Table 5.13 shows the GMSD's accuracy performance for each video distortion. The accuracy performance decreases for videos distorted by white noise, H.265, and MJPEG compression. But the accuracy performance decrease for H.265 Compression (1.37%) is lower than for MJPEG and

Table 5.15: Spearman Correlation Coefficient of PSNR on all databases.

Distortion	Video Resolution			
	768x432	640x360	480x270	384x216
Wavelet	0.7756	0.8004	0.8184	**0.8293**
H.264	0.7488	0.7743	0.7994	**0.8105**
H.265	0.8556	0.8782	0.8986	**0.9035**
MJPEG	**0.4672**	0.4296	0.2847	0.1694
MPEG2	0.7012	0.7289	0.7520	**0.7632**
Upscaling	0.6216	0.6402	0.6537	**0.6834**
Packet Loss	0.6591	0.6915	0.7115	**0.7187**
White Noise	0.9048	0.9060	**0.9079**	0.8947

white noise (16.54% and 8.21%). Even with the loss in accuracy performance for videos distorted by H.265 compression, the accuracy performance is still good.

For SSIM and PSNR when the resolution is reduced, the accuracy performance improves for almost all video distortions, except for videos distorted by white noise and MJPEG. Nevertheless, the accuracy performance of these methods are the worst among all methods, which is probably because these methods are actually image quality assessment methods.

If we reduce the resolutions of videos distorted by white noise and MJPEG compression the accuracy performances of the all video quality methods decrease. For videos distorted by packet loss, the accuracy performances of ViS3, STRRED and SSTS-GMSD decrease. For this reason, we propose to use the two classification methods to identify videos distorted by packet loss, MPJEG compression and white noise, and reduce the resolution of these videos to 768x432. The videos identified as distorted by the other types of distortions should have their spatial resolution reduced to 384x216.

5.2.1 Spatial Difference Classifier Results

The first classifier, based on Spatial Difference, divided the videos into two groups: videos distorted by packet loss, MPJEG compression and white noise; and videos distorted by upscaling, H.264, MPEG-2, H.265 and wavelet compressions. Table 5.16 shows the accuracy performance of each method using the spatial activity difference classifier. For comparison purposes, we also show the accuracy performance using: (1) the original resolution, (2) all videos in 384x216 and (3) an ideal classification. For ideal classification, we label all videos according to its distortion and then reduce the spatial resolution to 384x216 (MJPEG, packet loss, white noise) or 768x432 (others distortions). We notice that the accuracy performance is better than using the original resolution for all databases, except for STRRED in LIVE and CSIQ. Therefore, if we reduce the video resolution according to its distortion, the video quality assessment method accuracy performance can be improved.

Then, we analyze the accuracy performance using the classifier. We notice this approach is better than using the videos in the original resolution for CSIQ, IVP and MCL-V Databases. For

Table 5.16: Spearman Correlation Coefficient of all methods using Spatial Activity Difference Classifier.

Metric	Type of Classification	LIVE	CSIQ	IVP	MCL-V	All
ViS3	Original Resolution	**0.8168**	0.8325	0.7948	0.6361	0.7466
	All Videos in 384x216	0.7595	0.7897	**0.8912**	0.7674	0.8030
	Ideal Classification	0.7964	**0.8646**	0.8905	0.7674	**0.8381**
	SA Difference Classification	0.7610	0.8632	0.8809	**0.7727**	0.8294
STRRED	Original Resolution	**0.8007**	0.8129	**0.7374**	0.7433	0.6961
	All Videos in 384x216	0.7570	0.7424	0.7144	**0.8177**	0.7774
	Ideal Classification	0.7954	**0.8460**	0.7356	**0.8119**	**0.8176**
	SA Difference Classification	0.7363	0.7622	0.6774	0.7938	0.7698
SSTS-GMSD	Original Resolution	**0.8389**	0.8415	0.7560	0.6855	0.7403
	All Videos in 384x216	0.7777	0.7458	0.8722	0.8000	0.7855
	Ideal Classification	0.7364	0.8281	0.8699	0.8000	0.8028
	SA Difference Classification	0.7720	**0.8491**	**0.8780**	**0.8011**	**0.8250**
GMSD	Original Resolution	0.7262	0.8409	0.6924	0.6449	0.6781
	All Videos in 384x216	**0.7264**	0.7729	0.8672	0.7708	0.7874
	Ideal Classification	0.6787	0.8475	0.8611	0.7708	0.8020
	SA Difference Classification	0.7217	**0.8603**	**0.8708**	**0.7718**	**0.8225**
SSIM	Original Resolution	0.5251	0.5794	0.3560	0.4018	0.4833
	All Videos in 384x216	**0.7002**	0.7241	0.6965	**0.7105**	**0.7333**
	Ideal Classification	0.6567	0.6728	**0.7624**	**0.7105**	0.7217
	SA Difference Classification	0.6849	**0.7373**	0.6775	0.6613	0.7324
PSNR	Original Resolution	0.5233	0.5787	0.6566	0.4640	0.5795
	All Videos in 384x216	**0.6285**	0.5628	0.8374	**0.6876**	0.6857
	Ideal Classification	0.6239	**0.6284**	**0.8663**	**0.6876**	**0.7132**
	SA Difference Classification	0.6158	0.5974	0.8039	0.5964	0.6879

LIVE Database, the accuracy performance using the classifier is worst than using the videos in the original resolution for ViS3, STRRED, SSTS-GMSD and GMSD. On the other hand, for SSIM and PSNR the accuracy performance is better using the classifier.

The overall accuracy performance, considering all databases, for all methods using the classifier is better than using the videos in their original resolution. In some cases, the accuracy performance using the classifier is better than using the ideal classification. Probably, because the video content has an influence in the quality method accuracy performance. The spatial activity difference classifier is better to identify video contents that are more sensitive to resolution reduction.

Finally, the accuracy performance for videos in 384x216 resolution, i.e. if we reduce resolutions, regardless the distortion, is worst than using the classifier, except for SSIM. Therefore, if we want to improve the running time of this video quality assessment method and maintain their accuracy performance, we have to identify the video distortion and reduce its resolution accordingly.

From the SCC of the video quality assessment methods, we know that the predicted scores obtained via the proposed framework are more correlated with subjective scores than predicted scores obtained without using the framework. However, it is important to evaluate if the difference

Table 5.17: T-Test results between the SCC of various video quality assessment methods for all databases. The video quality assessment methods evaluate the videos in their original spatial resolution (Ori), 384x216 resolution (Min), using the framework with an ideal classifier (Ideal), and with SA classifier (SA).

		ViS3				STRRED				SSTS-GMSD			
		Ori	Min	Ideal	SA	Ori	Min	Ideal	SA	Ori	Min	Ideal	SA
ViS3	Ori	0	-1	-1	-1	1	-1	-1	-1	0	-1	-1	-1
	Min	1	0	-1	-1	1	0	-1	-1	1	-1	-1	-1
	Ideal	1	1	0	1	1	1	1	1	1	1	-1	-1
	SA	1	1	-1	0	1	1	-1	1	1	-1	-1	-1
STRRED	Ori	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1
	Min	1	0	-1	-1	1	0	-1	-1	1	-1	-1	-1
	Ideal	1	1	-1	1	1	1	0	0	1	1	-1	-1
	SA	1	1	-1	-1	1	1	0	0	1	-1	-1	-1
SSTSGMSD	Ori	0	-1	-1	-1	1	-1	-1	-1	0	-1	-1	-1
	Min	1	1	-1	1	1	1	-1	1	1	0	-1	-1
	Ideal	1	1	1	1	1	1	1	1	1	1	0	1
	SA	1	1	1	1	1	1	1	1	1	1	-1	0
GSMD	Ori	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	Min	1	1	-1	-1	1	1	-1	-1	1	-1	-1	-1
	Ideal	1	1	1	1	1	1	1	1	1	1	-1	-1
	SA	1	1	1	1	1	1	1	1	1	1	-1	-1
SSIM	Ori	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	Min	1	-1	-1	-1	1	-1	-1	-1	1	-1	-1	-1
	Ideal	1	-1	-1	-1	1	-1	-1	-1	1	-1	-1	-1
	SA	0	-1	-1	-1	1	-1	-1	-1	1	-1	-1	-1
PSNR	Ori	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	Min	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1
	Ideal	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1
	SA	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1

		GMSD				SSIM				PSNR			
		Ori	Min	Ideal	SA	Ori	Min	Ideal	SA	Ori	Min	Ideal	SA
ViS3	Ori	1	-1	-1	-1	1	-1	-1	0	1	1	1	1
	Min	1	-1	-1	-1	1	1	1	1	1	1	1	1
	Ideal	1	1	-1	-1	1	1	1	1	1	1	1	1
	SA	1	1	-1	-1	1	1	1	1	1	1	1	1
STRRED	Ori	1	-1	-1	-1	1	-1	-1	-1	1	-1	-1	0
	Min	1	-1	-1	-1	1	1	1	1	1	1	1	1
	Ideal	1	1	-1	-1	1	1	1	1	1	1	1	1
	SA	1	1	-1	-1	1	1	1	1	1	1	1	1
SSTSGMSD	Ori	1	-1	-1	-1	1	-1	-1	-1	1	1	1	1
	Min	1	1	-1	-1	1	1	1	1	1	1	1	1
	Ideal	1	1	1	1	1	1	1	1	1	1	1	1
	SA	1	1	1	1	1	1	1	1	1	1	1	1
GSMD	Ori	0	-1	-1	-1	1	-1	-1	-1	1	-1	-1	-1
	Min	1	0	-1	-1	1	1	1	1	1	1	1	1
	Ideal	1	1	0	1	1	1	1	1	1	1	1	1
	SA	1	1	-1	0	1	1	1	1	1	1	1	1
SSIM	Ori	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1
	Min	1	-1	-1	-1	1	0	-1	1	1	1	1	1
	Ideal	1	-1	-1	-1	1	1	0	1	1	1	1	1
	SA	1	-1	-1	-1	1	-1	-1	0	1	1	1	1
PSNR	Ori	-1	-1	-1	-1	1	-1	-1	-1	0	-1	-1	-1
	Min	1	-1	-1	-1	1	-1	-1	-1	1	0	-1	1
	Ideal	1	-1	-1	-1	1	-1	-1	-1	1	1	0	1
	SA	1	-1	-1	-1	1	-1	-1	-1	1	-1	-1	0

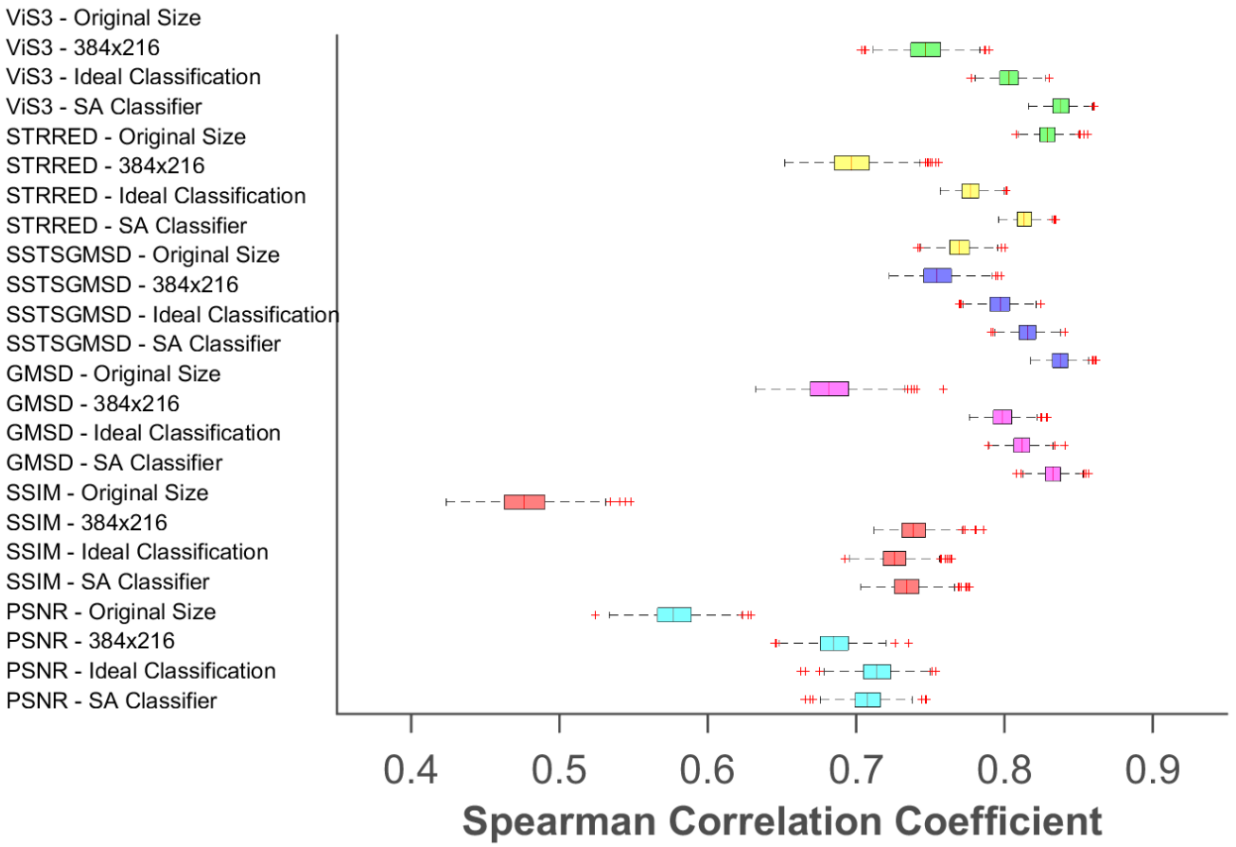


Figure 5.1: Box plot of SCC across 1000 simulations, where we ramdonly chose 80% of the reference videos and their distortions versions.

in results are statistically significant. To evaluate the statistical significance of the differences in accuracy performance with and without the framework, we performed a t-test on the SCC values obtained in 1,000 simulations [61]. In these simulations, we randomly chose 80% of the reference videos and their distorted versions for each simulation. The results of the t-test are shown in Table 5.17. The value "1" in the table indicates that the method in the row is statistically superior to the method in the column, and the value "-1" indicates that the method in the row is statistically inferior to the method in the column. Finally, the value "0" indicates that both methods are statistically indistinguishable. Notice that for the overall accuracy performance, considering all databases, the improvement using the framework with the spatial activity classifier is statistically significant. Figure 5.1 shows the box plot of the SCC across the 1000 simulations.

As said in the beginning of this chapter, the objective of reducing the video resolution is to reduce the running time of the video quality assessment method. The proposed classifier has the

Table 5.18: Average running time for performing an objective quality assessment with and without spatial activity classifier (in seconds).

	ViS3	STRRED	SSTS-GMSD	SSIM	GMSD	PNSR
Original Resolution	639.135	215.973	13.557	13.108	5.951	5.033
SA Difference Classifier	223.784	94.384	8.911	11.221	5.330	5.029

goal of identifying which videos can have its resolution reduced, avoiding a decrease in accuracy performance of the video quality assessment method. As mentioned in Chapter 4, we proposed a methodology to evaluate the video quality, composed of three stages. The first stage consist of identifying video distortion. The second stage reduces the video resolution to the smallest possible resolution, according to the video distortion. The third stage adjusts the predicted quality score, according to the video resolution. From the results in Table 5.16, we know that using the framework is a good way to maintain the accuracy performance of a quality assessment method. But, if we want to improve the runtime performance of video quality assessment methods, the running time of the classifier plus the running time of these methods has to be smaller than the running time of the video quality assessment method using the video in its original spatial resolution. Table 5.18 presents the average running time of the video quality assessment methods ran on videos in their original resolution and on videos with the resolution reduced, using the proposed methodology with the spatial difference classifier. From these results, we notice that, even adding the time spent to identify the video resolution and resizing the video, the proposed methodology is faster than the original video quality assessment method. When image quality assessment methods (GMSD,SSIM and PSNR) are used to measure video quality, the improvement in running time is small (SSIM is 1.17x faster, GMSD is 1.11x faster). Nevertheless, for

Table 5.19: Median Spearman Correlation Coefficient of all methods using SSEQ Distortion Classifier in 1000 simulations.

Metric	Type of Classification	LIVE	CSIQ	IVP	MCL-V	All
ViS3	Original Resolution	**0.8499**	0.8574	0.8470	0.5882	0.7532
	All Videos with 384x216	0.7740	0.8085	**0.9014**	**0.6765**	0.7953
	Ideal Classification	0.8202	**0.8736**	0.8998	**0.6765**	**0.8294**
	Predicted Classification	0.7882	0.8521	0.8946	**0.6765**	0.8164
STRRED	Original Resolution	**0.8263**	0.8093	0.7583	**0.9941**	0.7061
	All Videos with 384x216	0.7873	0.7565	**0.7931**	0.9853	0.7957
	Ideal Classification	0.7941	**0.8331**	0.7748	0.9853	**0.8263**
	Predicted Classification	0.7931	0.8055	0.7871	0.9853	0.8136
SSTS-GMSD	Original Resolution	**0.8625**	0.8389	0.7600	0.7529	0.7526
	All Videos with 384x216	0.8327	0.7961	**0.9146**	**0.9118**	0.8230
	Ideal Classification	0.8411	**0.8741**	0.8971	**0.9118**	**0.8501**
	Predicted Classification	0.8314	0.8565	0.9078	**0.9118**	0.8430
GMSD	Original Resolution	**0.7967**	0.8636	0.6705	0.7529	0.6757
	All Videos with384x216	0.7784	0.7761	**0.8757**	**0.9118**	0.8045
	Ideal Classification	0.7850	**0.8862**	0.8703	**0.9118**	**0.8419**
	Predicted Classification	0.7780	0.8644	0.8706	**0.9118**	0.8347
SSIM	Original Resolution	0.6535	0.6039	0.4105	0.4618	0.4877
	All Videos with384x216	**0.8167**	0.7426	0.7597	**0.9618**	0.7715
	Ideal Classification	0.7375	**0.7764**	**0.8331**	**0.9618**	**0.7906**
	Predicted Classification	0.7595	0.7282	0.7477	**0.9618**	0.7585
PSNR	Original Resolution	0.7112	0.6162	0.7187	0.5118	0.6052
	All Videos with384x216	**0.7517**	0.5887	0.8238	**0.9294**	0.7204
	Ideal Classification	0.6934	**0.6489**	**0.8656**	**0.9294**	**0.7402**
	Predicted Classification	0.6211	0.6108	0.8446	**0.9294**	0.7028

more complex video quality assessment methods, the improvement in time is considerable (ViS3 is 2.85x faster, STRRED is 2.29x faster and SSTS-GMSD is 1.51x faster).

5.2.2 SSEQ Classifier Results

The SSEQ Classifier uses a machine learning technique to classify the videos according to their distortion. The videos in all databases are divided in training and testing groups. 80% of the reference videos and their associated distorted versions were used for training and 20% of the reference videos and their associated distorted versions were used for testing. We use a C-SVM with linear kernel, as it was done in the original SSEQ algorithm for images. We combine videos in all databases and divided the whole set in training and testing groups. The training group consists of 472 to 480 video sequences, from which 120 video sequences are from LIVE Database, 180 video sequences are from CSIQ Database, 72 video sequences are from MCL-V Database and 100 to 108 video sequences are from IVP. Some reference video sequences in IVP Database do not have Packet Loss video sequences associated, that is why the number of video sequences in the train sets varies. The test set consist of 110 to 118 video sequences, from which 30 video sequences are from LIVE Database, 36 video sequences are from CSIQ Database, 24 video sequences are from MCL-V Database and 20 to 28 video sequences are from IVP. The training and testing sets do not share test sequences corresponding to the same content (original).

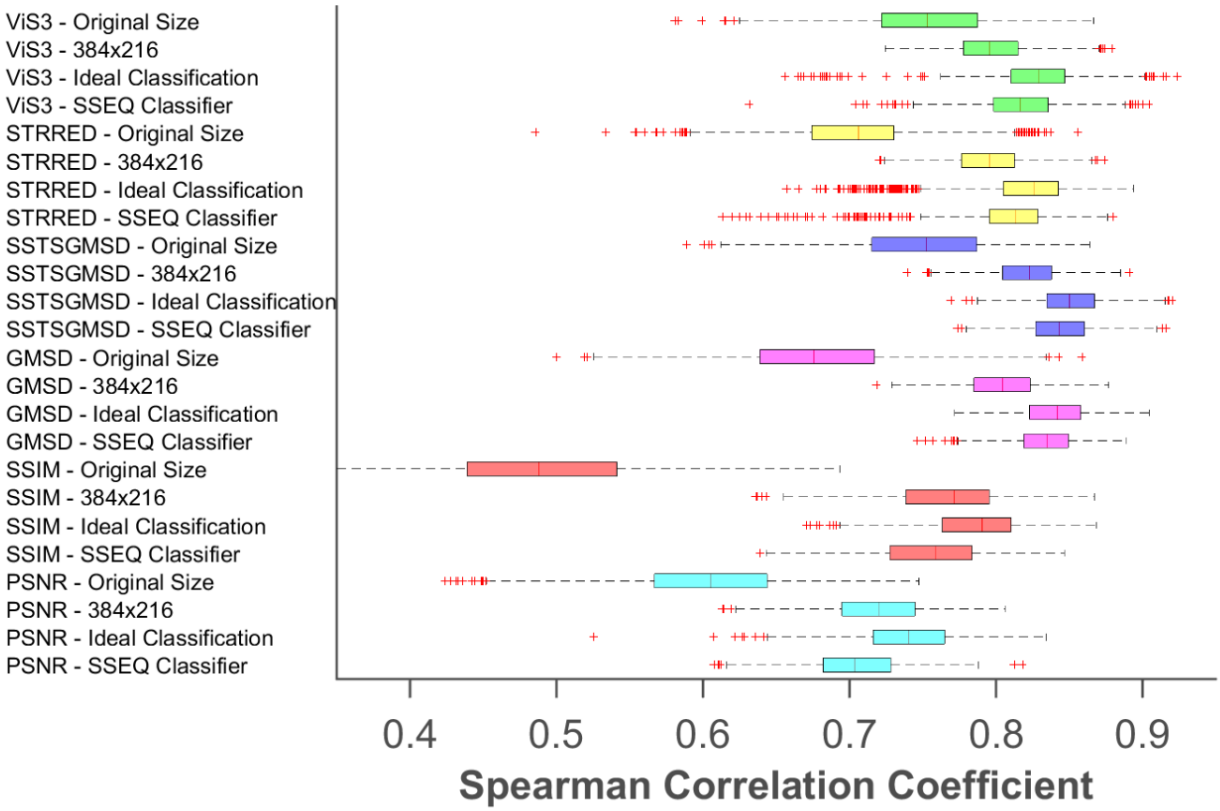


Figure 5.2: Box plot of SCC across 1000 test simulations, where we ramdonly chose 20% of the reference videos and their distortions versions.

Table 5.20: T-Test results between the SCC of various video quality assessment methods for all databases. The video quality assessment methods evaluate the videos in their original spatial resolution (Ori), 384x216 resolution (Min), using the framework with an ideal classifier (Ideal), and with SSEQ classifier (SSEQ).

		ViS3				STRRED				STS-GMSD			
		Ori	Min	Ideal	SSEQ	Ori	Min	Ideal	SSEQ	Ori	Min	Ideal	SSEQ
ViS3	Ori	0	-1	-1	-1	1	-1	-1	-1	0	-1	-1	-1
	Min	1	0	-1	-1	1	0	-1	-1	1	-1	-1	-1
	Ideal	1	1	0	1	1	1	1	1	1	1	-1	-1
	SSEQ	1	1	-1	0	1	1	-1	1	1	-1	-1	-1
STRRED	Ori	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1
	Min	1	0	-1	-1	1	0	-1	-1	1	-1	-1	-1
	Ideal	1	1	-1	1	1	1	0	0	1	1	-1	-1
	SSEQ	1	1	-1	-1	1	1	0	0	1	-1	-1	-1
SSTSGMSD	Ori	0	-1	-1	-1	1	-1	-1	-1	0	-1	-1	-1
	Min	1	1	-1	1	1	1	-1	1	1	0	-1	-1
	Ideal	1	1	1	1	1	1	1	1	1	1	0	1
	SSEQ	1	1	1	1	1	1	1	1	1	1	-1	0
GSMD	Ori	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	Min	1	1	-1	-1	1	1	-1	-1	1	-1	-1	-1
	Ideal	1	1	1	1	1	1	1	1	1	1	-1	-1
	SSEQ	1	1	1	1	1	1	1	1	1	1	-1	-1
SSIM	Ori	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	Min	1	-1	-1	-1	1	-1	-1	-1	1	-1	-1	-1
	Ideal	1	-1	-1	-1	1	-1	-1	-1	1	-1	-1	-1
	SSEQ	0	-1	-1	-1	1	-1	-1	-1	1	-1	-1	-1
PSNR	Ori	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	Min	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1
	Ideal	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1
	SSEQ	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1

		GMSD				SSIM				PSNR			
		Ori	Min	Ideal	SSEQ	Ori	Min	Ideal	SSEQ	Ori	Min	Ideal	SSEQ
ViS3	Ori	1	-1	-1	-1	1	-1	-1	0	1	1	1	1
	Min	1	-1	-1	-1	1	1	1	1	1	1	1	1
	Ideal	1	1	-1	-1	1	1	1	1	1	1	1	1
	SSEQ	1	1	-1	-1	1	1	1	1	1	1	1	1
STRRED	Ori	1	-1	-1	-1	1	-1	-1	-1	1	-1	-1	0
	Min	1	-1	-1	-1	1	1	1	1	1	1	1	1
	Ideal	1	1	-1	-1	1	1	1	1	1	1	1	1
	SSEQ	1	1	-1	-1	1	1	1	1	1	1	1	1
SSTSGMSD	Ori	1	-1	-1	-1	1	-1	-1	-1	1	1	1	1
	Min	1	1	-1	-1	1	1	1	1	1	1	1	1
	Ideal	1	1	1	1	1	1	1	1	1	1	1	1
	SSEQ	1	1	1	1	1	1	1	1	1	1	1	1
GSMD	Ori	0	-1	-1	-1	1	-1	-1	-1	1	-1	-1	-1
	Min	1	0	-1	-1	1	1	1	1	1	1	1	1
	Ideal	1	1	0	1	1	1	1	1	1	1	1	1
	SSEQ	1	1	-1	0	1	1	1	1	1	1	1	1
SSIM	Ori	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1
	Min	1	-1	-1	-1	1	0	-1	1	1	1	1	1
	Ideal	1	-1	-1	-1	1	1	0	1	1	1	1	1
	SSEQ	1	-1	-1	-1	1	-1	-1	0	1	1	1	1
PSNR	Ori	-1	-1	-1	-1	1	-1	-1	-1	0	-1	-1	-1
	Min	1	-1	-1	-1	1	-1	-1	-1	1	0	-1	1
	Ideal	1	-1	-1	-1	1	-1	-1	-1	1	1	0	1
	SSEQ	1	-1	-1	-1	1	-1	-1	-1	1	-1	-1	0

Table 5.21: Average running time for performing an objective quality assessment with and without SSEQ classifier (in seconds).

	ViS3	STRRED	SSTS-GMSD	SSIM	GMSD	PNSR
Original Resolution	647.5480	219.9712	13.7975	13.6208	6.1423	5.1625
SA Difference Classification	278.4741	157.7950	73.9999	76.2293	70.5274	70.2187

In summary, for 1,000 simulations, the databases are randomly divided in traning and testing groups, respecting content non overlap. Table 5.19 shows the median of the SCC values obtained for these 1,000 simulations.

We compare the results obtained with the SSEQ classifier with the results obtained with the videos in the original resolution, the videos in 384x216 resolution, and using an ideal classification. Figure 5.2 shows the box plot of SCC across 1,000 simulations for the overall accuracy performance combining all databases. From these results, we notice that using the ideal classification generates the best results for all methods.

The gain in accuracy performance using the ideal classification is 10.14% for ViS3, 17.30% for STRRED, 13.32% for SSTS-GMSD, 25.01% for GMSD, 60.85% for SSIM and 22.17% for PSNR. SSEQ classifier gives the overall second best results for most methods, with the exception of SSIM and PSNR. The gain in accuracy performance using the SSEQ classifier is 8.37% for ViS3, 15.70% for STRRED, 12.64% for SSTS-GMSD, 23.94% for GMSD, 53.74% for SSIM and 15.96% for PSNR. The highest gains correspond to the MCL-V and IVP databases, which have the videos with a higher resolution. Also, the gain in accuracy performance of image quality methods, is higher than the other methods. And, when using the SSEQ classifier, SSTS-GMSD has the second best accuracy performance of all methods.

The statistical significance of the differences in accuracy performance was made using the t-test across the 1,000 test simulations. Tables 5.20 shows the results of the t-test. The improvement when using the framework with the SSEQ classifier is statistically significant. Notice that the results of STRRED using the ideal classifier and of the SSEQ classifier are equivalent, even though the SCC of STRRED with ideal classifier is greater than the SCC of STRRED with SSEQ classifier. It is worth noticing that all methods, excet for PSNR, have a better accuracy performance using the framework than the ViS3 accuracy performance on videos with their original resolution.

Table 5.21 presents the average running time for video quality assessment methods for videos in their original resolution and with the proposed methodology using the SSEQ classifier. Unfortunately, the SSEQ classifier is slower than the SA Classifier, because it extracts more features from the video. Only the running time of ViS3 and STRRED is improved when using the SSEQ classifier (ViS3 is 2.32x faster and STRRED is 1.39x faster).

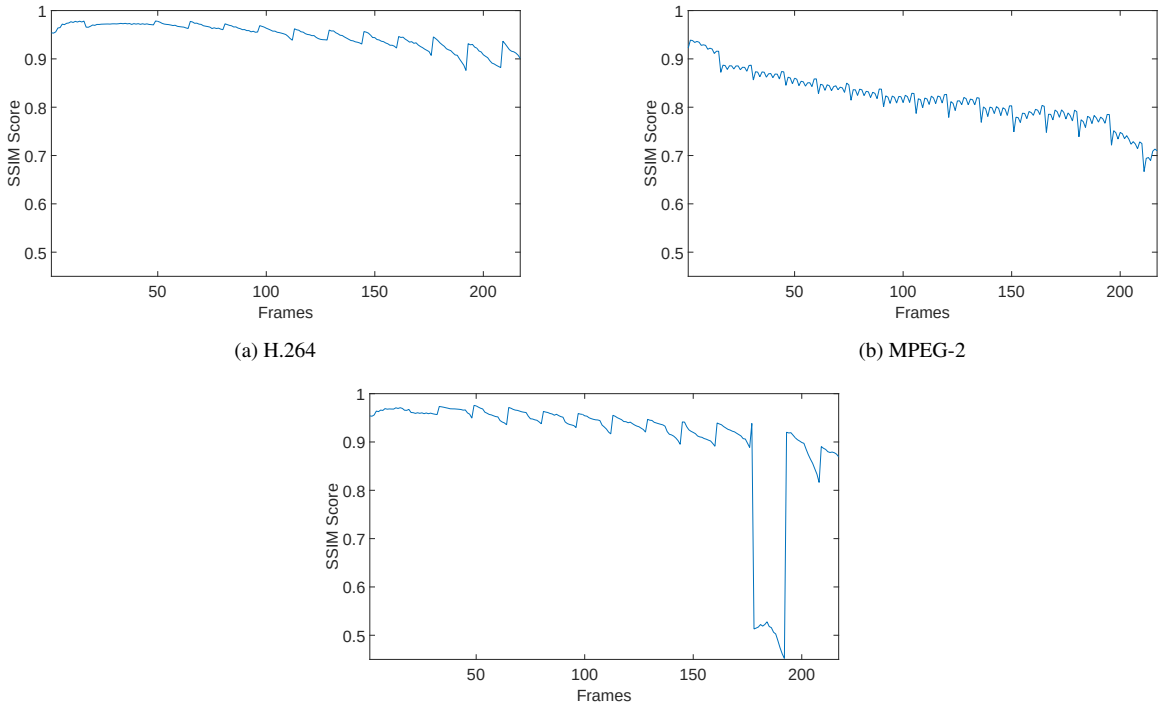
Table 5.22: Spearman Coefficient Correlation of the methods when varying the video FPS.

Métrica	Banco de Dados	FPS				
		Original	24	16	8	4
ViS3	LIVE	0.8168	**0.8191**	0.8108	0.8021	0.7586
	CSIQ	0.8325	**0.8379**	0.8347	0.8040	0.7573
	IVP	0.7948	0.7931	**0.7998**	0.7439	0.7466
	MCL-V	0.6361	0.6500	**0.6933**	0.6693	-
STRRED	LIVE	0.8007	**0.8074**	0.8002	0.7888	0.7773
	CSIQ	**0.8129**	0.7995	0.7964	0.7799	0.7617
	IVP	0.7374	0.7396	**0.7475**	0.7403	0.7393
	MCL-V	**0.7433**	0.7431	0.7381	0.7281	0.7222
SSTS-GMSD	LIVE	**0.8389**	0.8262	0.8182	0.7967	0.7813
	CSIQ	0.8415	**0.8492**	0.8474	0.8430	0.8299
	IVP	0.7560	0.7490	0.7553	0.7662	**0.7863**
	MCL-V	0.6855	0.6881	0.6937	0.6932	**0.7008**
GMSD	LIVE	0.7262	0.7311	0.7108	**0.7358**	0.6900
	CSIQ	**0.8409**	0.8391	0.8390	0.8383	0.8364
	IVP	**0.6924**	0.6839	0.6785	0.6894	0.6691
	MCL-V	0.6449	0.6490	0.6477	0.6514	**0.6608**
SSIM	LIVE	0.5251	0.5210	0.5010	**0.5400**	0.5137
	CSIQ	0.5794	0.5793	0.5786	**0.5852**	0.5765
	IVP	0.3560	0.3556	0.3493	**0.3568**	0.3404
	MCL-V	0.4018	0.4009	0.4050	0.3976	**0.4066**
PSNR	LIVE	0.5251	0.5096	0.4921	0.5348	**0.5376**
	CSIQ	0.5794	0.5791	0.5761	0.5842	**0.5847**
	IVP	**0.6566**	0.6469	0.6376	0.6501	0.6274
	MCL-V	0.4640	0.4683	0.4659	0.4693	**0.4771**

5.2.3 Influence of Frame Rate in Video Quality Metrics

Other way to improve the runtime performance of video quality assessment methods is to reduce the frame rate of the videos. Table 5.22 shows the SCC of the video quality assessment methods when reducing the video frame rate to 24, 16, 8 and 4 FPS. There are no ViS3 results for videos with 4 FPS in MCL-V database, because these videos do not have enough temporal information required by ViS3. From these results, we notice that ViS3 and STRRED accuracy performances are better with a higher FPS. We believe this happens because these algorithms analyze the temporal information of the video to measure the its quality. For example, ViS3 uses motion vectors to give weight the to specific areas in the spatial distortion map. Therefore, in this case, by reducing the temporal resolution we might introduce temporal transitions to the video. STRRED, on the other hand, analyses the differences among neighboring frames in the video. Therefore, if there is a large time interval between to two consecutive frames, there will be a large temporal difference between them. Nevertheless, for these two video quality assessment methods we can reduce the FPS down to 16, without greatly affecting the accuracy performance.

The accuracy performance of SSTS-GMSD is higher for higher FPS, both for LIVE and CSIQ databases, and for lower FPS, both for IVP and MCL-V databases. We believe this happens



(a) H.264

(b) MPEG-2

(c) Packet Loss

Figure 5.3: SSIM score of 'Blue Sky' video distorted by H.264 Compression, MPEG-2 Compression and Packet Loss.

because SSTS-GMSD also performs a temporal analysis. At the same time, the CSIQ, LIVE and IVP Databases have videos distorted by packet loss, which is a temporal distortion. The videos in the CSIQ and LIVE databases are in 768x432 and 832x480 resolution, respectively, and 24, 25, 30, 50 and 60 FPS, while videos in the IVP and MCL-V databases are in 1088x1920 and 1080x1920 resolution, respectively, and 25 and 30 FPS. Therefore, when we discard frames from videos in the CSIQ and LIVE databases, a lot more information is being discarded.

For GMSD and SSIM, using videos sequences in 8 FPS is a reasonable trade-off between accuracy performance and runtime performance. The GMSD accuracy performance for the LIVE and MCL-V databases are better with videos sequence in 8 FPS than in the original FPS. For CSIQ and IVPL, there is a loss in accuracy performance is 0.3% and 0.4%, respectively, when we compare videos in the original FPS with videos in 8 FPS. The SSIM accuracy performance is best for videos in 8 FPS for LIVE, CSIQ and IVP databases. For the IVP database, the loss in accuracy performance is 1.04% when we compare videos in the original FPS with videos in 8 FPS. The best accuracy performance for PSNR is for videos in 4 FPS for LIVE, CSIQ and MCL-V database. For IVP database, the PSNR accuracy performance with videos in 4 FPS is 0.99% worst than with videos in the original FPS.

The GMSD, SSIM and PSNR image quality assessment methods can be used to measure video quality by using them to measure the quality of each frame and averaging the frame quality scores to obtain the overall video quality score. Figures 5.3 (a), (b), and (c) show the SSIM frame scores for a video sequence distorted by H.264 Compression, MPEG-2 compression and Packet

Table 5.23: Spearman Coefficient Correlation of the metrics when varying the video FPS for each video distortion.

Metric	Distortion	FPS				
		Original	24	16	8	4
ViS3	Wavelet	**0.6537**	0.6329	0.6345	0.5772	0.5132
	H.264	0.7413	0.7479	**0.7486**	0.7481	-
	H.265	0.9174	**0.9236**	0.9151	0.8878	0.8124
	MJPEG	0.7349	**0.7671**	0.7560	0.6533	0.5613
	MPEG2	0.5278	**0.5308**	0.5239	0.5028	0.5140
	Upscaling	0.6890	**0.6927**	0.7369	0.7200	-
	Packet Loss	**0.8163**	0.8128	0.7993	0.7643	0.7335
	White Noise	**0.9202**	0.9156	0.9125	0.8826	0.8669
STRRED	Wavelet	**0.7825**	0.7608	0.7579	0.7419	0.7303
	H.264	0.7718	**0.7790**	0.7768	0.7787	0.7605
	H.265	0.9135	0.9187	0.9230	**0.9310**	0.9212
	MJPEG	**0.7290**	0.7223	0.7145	0.6857	0.6752
	MPEG2	**0.6772**	0.6615	0.6530	0.6289	0.6262
	Upscaling	0.7040	**0.7067**	0.7006	0.6954	0.6902
	Packet Loss	0.8016	0.8067	**0.8068**	0.8014	0.7792
	White Noise	0.9305	**0.9377**	0.9179	0.9192	0.9292
SSTS-GMSD	Wavelet	**0.5981**	0.5915	0.5783	0.5546	0.5345
	H.264	**0.8025**	0.7936	0.7921	0.7802	0.7861
	H.265	0.9287	0.9359	**0.9418**	0.9375	0.9261
	MJPEG	0.8803	**0.8821**	0.8736	0.8690	0.8396
	MPEG2	**0.6272**	0.6136	0.6062	0.6053	0.6181
	Upscaling	0.6795	0.6806	0.6854	0.6936	**0.6945**
	Packet Loss	**0.7899**	0.7842	0.7797	0.7708	0.7516
	White Noise	0.8819	**0.8860**	0.8811	0.8682	0.8456
GMSD	Wavelet	0.5812	0.5798	**0.5818**	0.5805	0.5636
	H.264	0.7531	0.7517	0.7451	0.7640	**0.7787**
	H.265	0.9418	**0.9441**	0.9398	0.9416	0.9369
	MJPEG	0.8842	0.8842	0.8847	**0.8893**	0.8744
	MPEG2	0.5501	**0.5634**	0.5371	0.5435	0.5569
	Upscaling	0.6376	0.6378	0.6371	0.6400	**0.6549**
	Packet Loss	0.7770	**0.7795**	0.7708	0.7657	0.7172
	White Noise	0.9094	0.9102	0.9102	**0.9130**	0.9066
SSIM	Wavelet	0.6130	0.6141	**0.6167**	0.6106	0.6029
	H.264	0.6257	0.6228	0.6223	0.6333	**0.6464**
	H.265	0.8136	0.8139	0.8100	0.8124	**0.8172**
	MJPEG	0.7969	0.7969	0.7969	**0.8054**	0.7974
	MPEG2	0.4814	0.4854	0.4825	0.4791	**0.5224**
	Upscaling	0.4400	0.4408	**0.4467**	0.4406	0.4425
	Packet Loss	0.4485	0.4504	0.4433	**0.4541**	0.4166
	White Noise	**0.9300**	**0.9300**	0.9282	0.9274	**0.9300**
PSNR	Wavelet	0.6696	**0.6716**	0.6697	0.6711	0.6623
	H.264	0.6631	0.6616	0.6554	0.6714	**0.6866**
	H.265	0.7846	0.7938	0.7701	**0.8015**	0.7969
	MJPEG	0.5086	0.5112	0.5086	0.5086	**0.5341**
	MPEG2	0.5371	0.5389	0.5282	0.5465	**0.5475**
	Upscaling	0.4925	0.5015	0.4996	0.5045	**0.5151**
	Packet Loss	**0.6509**	0.6489	0.6401	0.6373	0.6252
	White Noise	0.9063	0.9030	0.9063	**0.9071**	0.9009

Loss, respectively. Notice that the neighboring frames have similar SSIM frame scores for videos distorted by compression. SSIM frame scores for videos distorted by Packet Loss are very different when a packet loss occurs, but similar for the rest of the frames. For temporal distortions, like packet loss, discarding frames from the video gives reduces the accuracy of the predicted score more than reducing the spatial resolution. So, unless the video has temporal distortions, it is fine to discard some frames when using GMSD, SSIM and PSNR. Table 5.23 shows that, in most cases, the methods can perform best for packet loss when using videos in the original temporal resolution or with 24 FPS.

5.2.4 Discussion

One of open challenges in image/video quality assessment is the computational time performance [12, 13]. We notice from Table 5.18 that ViS3 and STRRED takes more than three minutes, on average, to execute. Considering that all videos are, approximately, 10 seconds long, these methods cannot be used in real time or practical applications.

A simple method to reduce the computational time is to reduce the video spatial resolution, but, sometimes, this strategy can decrease the accuracy performance of the methods, as seen in Tables 5.2 5.3. However, for IVP and MCL-V databases the accuracy performance improves when the video spatial resolution is reduced. Each database has its set of video distortions and from the results shown in Tables 5.6- 5.9, we notice that the methods have a different accuracy performance for each distortion when the video spatial resolution is reduced. We create two distortion classifiers to identify distortions that video quality method has better accuracy performance at 768x432 and distortions that video quality method has better accuracy performance at 384x216. Tables 5.16 and 5.19 show the accuracy performance using these classifiers. In both cases the overall accuracy performance using the classifier is better than when the videos are in their original resolution. We notice that for image quality assessment methods (PSNR, SSIM and GMSD) the accuracy performance improvement is greater than for other methods. We have shown that it is important to have a distortion classifier, given that the accuracy performance when using all videos in 384x216 is worse than using the classifier.

Figures 5.4 and 5.5 shows the runtime performance of the spatial activity classifier and the SSEQ classifier. The spatial activity classifier is a very fast method, being faster to use the video quality assessment method with the classifier than the method by itself. From the results of the method accuracy performance, we noticed that the framework with the spatial activity classifier improves both the computational and accuracy performances. The SSEQ classifier is slower than the spatial activity classifier, since it extracts 12 features from the video, and the spatial activity classifier extracts only one feature. But, the SSEQ classifier is able to improve the runtime performance of the ViS3 and STRRED.

Finally, we analyze how temporal resolution reduction affects the video quality accuracy performance. Video quality assessment method that analyze temporal information had their accuracy performance reduced. However image quality assessment method are not very affected by tem-

poral resolution reduction.

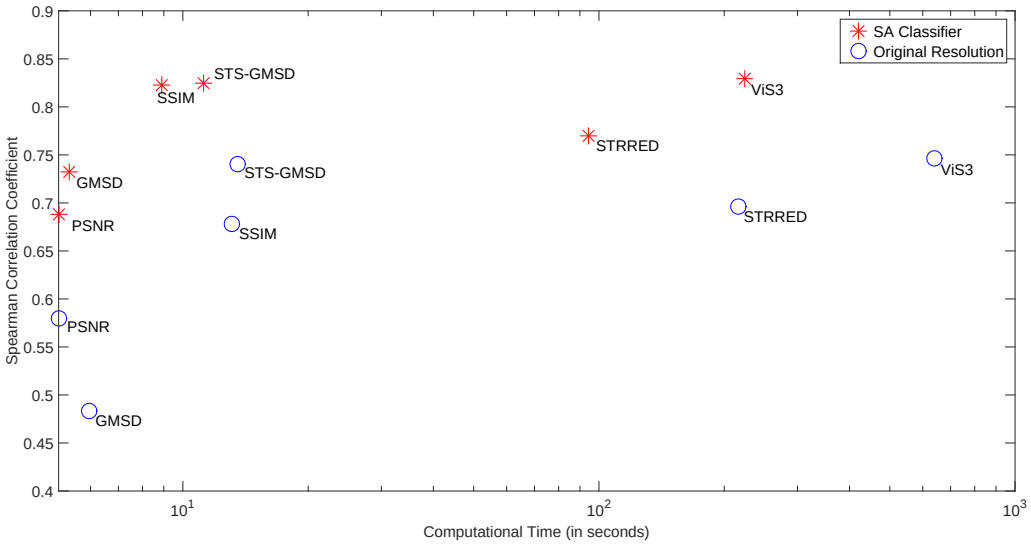


Figure 5.4: Spearman correlation coefficient versus computational time (in log space). Comparing the accuracy performance when using the Spatial Activity classifier and the video in their original resolution.

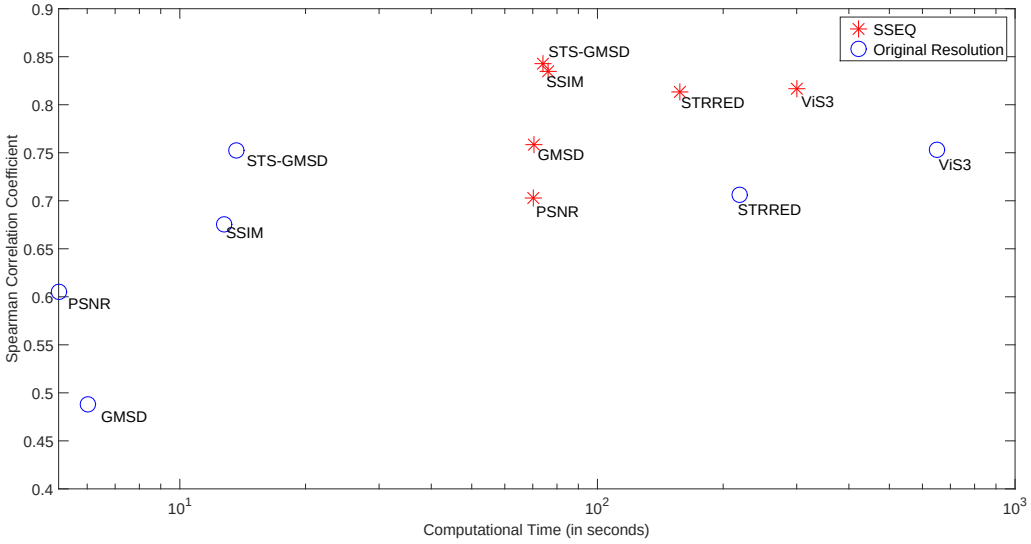


Figure 5.5: Spearman correlation coefficient versus computational time (in log space). Comparing the accuracy performance when using the SSEQ classifier and the video in their original resolution.

6 CONCLUSIONS

Over the decades, video quality assessment methods have been design with the goal of estimating video quality. Unfortunately, these methods have become more complex and most methods cannot be used in real time scenarios or any practical application. For example, ViS3 and STRRED take minutes to estimate the quality of a video 10 seconds long. Therefore, one of the challenges of video quality assessment methods is how to improve the running time performance [12].

One of the fastest and simplest methods to improve the running time performance is reducing the spatial resolution. So, in this work, we analyzed the accuracy performance of video quality assessment methods when the spatial resolution is reduced. Six video quality assessment methods are used in this work: the PSNR, SSIM, GMSD, SSTS-GMSD, STRRED and ViS3. PSNR, SSIM, GMSD are image quality assessment methods, which can be used to compute video quality by averaging the predicted scores obtained for each frame. SSTS-GMSD is a video quality assessment method based on the GMSD; STRRED is a video quality assessment method based on the entropic difference between the wavelet coefficients; and ViS3 is a video quality assessment method that computes the spatial distortions and the spatiotemporal dissimilarity between the reference and distorted video. These methods were chosen because each of them evaluate different features from the video to calculated its quality.

The spatial resolution analysis shows that the method's accuracy performance is affected differently when the spatial resolution is reduced. Videos distorted by MJPEG compression, packet loss and white noise are the most sensitive to the reduction of spatial resolution.

In this work, we proposed a framework to improve the running time of video quality methods, without decreasing the video quality assessment methods accuracy performance. This framework contains four stages: identification of video distortion, spatial downsampling of the video according to the distortion, estimation of the video quality using objective video quality assessment methods, and adjustment of the predicted quality scores. For the first stage, we propose two classification methods to identify videos that are more sensitive to the spatial resolution reduction.

The first classification method is based on the spatial activity of the video. We notice that the videos distorted by MJPEG compression, packet loss and white noise have a higher spatial activity than its reference video. The overall results using this classification method are faster and better than using the only the video quality assessment method.

The second classification method is based on spatial an spectral entropies. It was originally developed as part of a image quality assessment method [50]. We adapted it to work as a video classification, using the average value of the frame features as the feature vector. This classification method uses a support vector machine to classify the videos. The results using this classification method have a better overall accuracy performance than using only the video quality

assessment method. But, it has a slow feature extraction process, and because of it, this method only improves the speed of the STRRED and ViS3. Therefore, using the framework can improve the running time of the video quality assessment method without affecting the accuracy performance. However, the classification method has to be fast and precise.

We also perform a temporal resolution analysis. This analysis shows that the video quality assessment methods, which use temporal analysis, have a loss in accuracy performance when the temporal resolution of videos is reduced. Independent of the video quality assessment method, the method's accuracy performance of videos distorted by packet loss have a loss in performance for lower video resolution. Because packet loss is a temporal distortion that appear in any frame of the video, when we reduce the temporal resolution the frames affected by packet loss can be discarded.

6.1 FUTURE WORKS

We studied the influence of the video distortion on the performance of video quality assessment methods when reducing the spatial resolution. When comparing the results for the spatial activity classifier and the ideal classification, we noticed that sometimes the spatial activity classifier gives better results. This means that it is not only the video distortion that influences the accuracy performance, when reducing the spatial resolution. Probably, the content of the videos also influences the method's performance. So, future works include analyzing the influence of the video content on the method's accuracy performance.

Also future work includes, performing an analysis of the influence of both spatial and temporal resolution reductions on the accuracy performance of video quality assessment method. Also, it would be important to find the optimal spatial and temporal resolutions, to improve the proposed framework. Although the temporal resolution is important for the methods that analyze temporal distortions, for GMSD, SSIM and PSNR a reduction of the temporal resolution does not affect the accuracy performance.

Finally, we could test our framework with a set of no-reference video quality assessment methods and use the SSEQ as the video classifier.

7 CONTRIBUTIONS

Our final objective is to create a framework, that is able to reduce the runtime performance of video quality assessment methods. To achieve this objective, we did several analysis and proposed two video classifier. These analysis and classifiers are also contributions of our work.

We analyze the accuracy performance of six video quality metrics in four video databases, when the spatial video resolution is reduced. We did an analysis combining all databases to identify which video distortions are more sensitive to spatial resolution reduction. And from that analysis we were able to proposed two simple and fast video classifiers, that identify videos more sensitive to spatial resolution reduction.

Also, we perform a temporal resolution analysis using the same video quality assessment methods and databases. From that analysis, we conclude that video quality assessment methods, which use temporal analysis are more affected by the temporal resolution reduction. And, we perform an analysis, combining all databases, to identify the video distortions more sensitive to temporal resolution reduction.

Finally, our major contribution is our proposed framework, which from our results is able to improve the runtime and accuracy performance of video quality assessment methods.

8 PUBLISHED WORKS

1. Freitas, Pedro Garcia, Welington YL Akamine, and Mylène CQ Farias. "No-Reference Image Quality Assessment Using Texture Information Banks," *5th Brazilian Conference on Intelligent Systems (BRACIS)*, Recife, Brazil, 2016, pp. 127-132.

2. Freitas, Pedro Garcia, Welington YL Akamine, and Mylène CQ Farias. "Blind Image Quality Assessment Using Multiscale Local Binary Patterns." *Journal of Imaging Science and Technology* 60.6 (2016): 60405-1.

3. Freitas, Pedro Garcia, Welington YL Akamine, and Mylene CQ Farias. "No-reference image quality assessment based on statistics of Local Ternary Pattern." *Quality of Multimedia Experience (QoMEX)*, 2016 Eighth International Conference on. IEEE, 2016.

Bibliography

1 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.

2 YAN, P.; MOU, X.; XUE, W. Video quality assessment via gradient magnitude similarity deviation of spatial and spatiotemporal slices. *Proc. SPIE*, v. 9411, p. 94110M–94110M–10, 2015. Disponível em: <http://dx.doi.org/10.1117/12.2083283>.

3 SIMONCELLI, E. P.; FREEMAN, W. T. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In: IEEE. *Image Processing, 1995. Proceedings., International Conference on*. [S.l.], 1995. v. 3, p. 444–447.

4 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>.

5 CISCO Visual Networking Index: Forecast and Methodology, 2014-2019 White Paper. 2015.

6 SESHADRINATHAN, K.; SOUNDARARAJAN, R.; BOVIK, A. C.; CORMACK, L. K. Study of subjective and objective quality assessment of video. *IEEE transactions on image processing*, IEEE, v. 19, n. 6, p. 1427–1441, 2010.

7 WANG, Z.; BOVIK, A. C. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, v. 26, n. 1, p. 98–117, Jan 2009. ISSN 1053-5888.

8 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.

9 VU, P. V.; VU, C. T.; CHANDLER, D. M. A spatiotemporal most-apparent-distortion model for video quality assessment. In: IEEE. *2011 18th IEEE International Conference on Image Processing*. [S.l.], 2011. p. 2505–2508.

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

11 SESHADRINATHAN, K.; BOVIK, A. C. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE transactions on image processing*, IEEE, v. 19, n. 2, p. 335–350, 2010.

12 CHANDLER, D. M. Seven challenges in image quality assessment: past, present, and future research. *ISRN Signal Processing*, Hindawi Publishing Corporation, v. 2013, 2013.

13 WANG, Z. Objective image quality assessment: Facing the real-world challenges. *Electronic Imaging*, Society for Imaging Science and Technology, v. 2016, n. 13, p. 1–6, 2016.

14 FARIAS, M. C. *Video quality metrics*. [S.l.]: INTECH Open Access Publisher, 2010.

15 SAYOOD, K. *Introduction to Data Compression*. [S.l.]: Morgan Kaufmann, 2012. (Morgan Kaufmann series in multimedia information and systems). ISBN 9780124157965.

16 AMER, A.; MITICHE, A.; DUBOIS, E. Reliable and fast structure-oriented video noise estimation. In: IEEE. *Image Processing. 2002. Proceedings. 2002 International Conference on*. [S.l.], 2002. v. 1, p. I–840.

17 RISSANEN, J. Modeling by shortest data description. *Automatica*, Elsevier, v. 14, n. 5, p. 465–471, 1978.

18 PENNEBAKER, W. B.; MITCHELL, J. L. *JPEG: Still image data compression standard*. [S.l.]: Springer Science & Business Media, 1992.

19 UNTERWEGER, A. Compression artifacts in modern video coding and state-of-the-art means of compensation. *Multimedia Networking and Coding*, IGI Global, p. 28, 2012.

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

21 TUDOR, P. Mpeg-2 video compression. *Electronics & communication engineering journal*, IET, v. 7, n. 6, p. 257–264, 1995.

22 WIEGAND, T.; SULLIVAN, G. J.; BJONTEGAARD, G.; LUTHRA, A. Overview of the h. 264/avc video coding standard. *IEEE Transactions on circuits and systems for video technology*, IEEE, v. 13, n. 7, p. 560–576, 2003.

23 KOH, C. C.; MITRA, S. K.; FOLEY, J. M.; HEYNDERICKX, I. E. Annoyance of individual artifacts in mpeg-2 compressed video and their relation to overall annoyance. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *Electronic Imaging 2005*. [S.l.], 2005. p. 595–606.

24 LIST, P.; JOCH, A.; LAINEMA, J.; BJONTEGAARD, G.; KARCZEWICZ, M. Adaptive deblocking filter. *IEEE transactions on circuits and systems for video technology*, v. 13, n. 7, p. 614–619, 2003.

25 BORER, T.; DAVIES, T. Dirac video compression using open technology. *BBC EBU technical review*, 2005.

26 SULLIVAN, G. J.; OHM, J.-R.; HAN, W.-J.; WIEGAND, T. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, IEEE, v. 22, n. 12, p. 1649–1668, 2012.

27 OHM, J.-R.; SULLIVAN, G. J.; SCHWARZ, H.; TAN, T. K.; WIEGAND, T. Comparison of the coding efficiency of video coding standards—including high efficiency video coding (hevc). *IEEE Transactions on Circuits and Systems for Video Technology*, IEEE, v. 22, n. 12, p. 1669–1684, 2012.

28 LIN, J. Y.; SONG, R.; WU, C.-H.; LIU, T.; WANG, H.; KUO, C.-C. J. Mcl-v: A streaming video quality assessment database. *Journal of Visual Communication and Image Representation*, Elsevier, v. 30, p. 1–9, 2015.

29 PARKER, J. A.; KENYON, R. V.; TROXEL, D. E. Comparison of interpolating methods for image resampling. *IEEE Transactions on medical imaging*, IEEE, v. 2, n. 1, p. 31–39, 1983.

30 BOULOS, F.; PARREIN, B.; CALLET, P. L.; HANDS, D. Perceptual effects of packet loss on h. 264/avc encoded videos. In: *Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM-09*. [S.l.: s.n.], 2009.

31 SESHADRINATHAN, K.; SOUNDARARAJAN, R.; BOVIK, A. C.; CORMACK, L. K. A subjective study to evaluate video quality assessment algorithms. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *IS&T/SPIE Electronic Imaging*. [S.l.], 2010. p. 75270H–75270H.

32 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.

33 KOZAMERNIK, F.; SUNNA, P.; WYCKENS, E.; PETTERSEN, D. I. Subjective quality of internet video codecs phase ii evaluations using samviq. *EBU technical Review*, v. 301, 2005.

34 ZHANG, F.; LI, S.; MA, L.; WONG, Y. C.; NGAN, K. N. *IVP subjective quality video database*. 2011. Disponível em: <http://ivp.ee.cuhk.edu.hk/research/database/subjective/>.

35 RECOMMENDATION, P. I.-T. Subjective video quality assessment methods for multimedia applications. 1999.

36 BT.500-11, I.-R. R. Methodology for the subjective assessment of the quality of television pictures. v. 2002. Disponível em: <https://www.itu.int/>.

37 WINKLER, S.; MOHANDAS, P. The evolution of video quality measurement: From psnr to hybrid metrics. *IEEE Transactions on Broadcasting*, v. 54, n. 3, p. 660–668, Sept 2008. ISSN 0018-9316.

38 LUBIN, J.; FIBUSH, D. *Sarnoff JND vision model*. [S.l.]: T1A1, 1997.

39 WINKLER, S. Perceptual distortion metric for digital color video. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *Electronic Imaging'99*. [S.l.], 1999. p. 175–184.

40 XUE, W.; ZHANG, L.; MOU, X.; BOVIK, A. C. Gradient magnitude similarity deviation: a highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, IEEE, v. 23, n. 2, p. 684–695, 2014.

41 LARSON, E. C.; CHANDLER, D. M. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, International Society for Optics and Photonics, v. 19, n. 1, p. 011006–011006, 2010.

42 XU, J.; YE, P.; LIU, Y.; DOERMANN, D. No-reference video quality assessment via feature learning. In: IEEE. *2014 IEEE International Conference on Image Processing (ICIP)*. [S.l.], 2014. p. 491–495.

43 SAAD, M. A.; BOVIK, A. C.; CHARRIER, C. Blind prediction of natural video quality. *IEEE Transactions on Image Processing*, IEEE, v. 23, n. 3, p. 1352–1365, 2014.

44 SOUNDARARAJAN, R.; BOVIK, A. C. Video quality assessment by reduced reference spatio-temporal entropic differencing. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 23, n. 4, p. 684–694, April 2013. ISSN 1051-8215.

45 SOUNDARARAJAN, R.; BOVIK, A. C. Rred indices: Reduced reference entropic differencing for image quality assessment. *IEEE Transactions on Image Processing*, IEEE, v. 21, n. 2, p. 517–526, 2012.

46 WAINWRIGHT, M. J.; SIMONCELLI, E. P. Scale mixtures of gaussians and the statistics of natural images. p. 855–861, 1999.

47 LUCAS, B. D.; KANADE, T. et al. An iterative image registration technique with an application to stereo vision. In: *IJCAI*. [S.l.: s.n.], 1981. v. 81, n. 1, p. 674–679.

48 FENIMORE, C.; LIBERT, J.; WOLF, S. Perceptual effects of noise in digital video compression. *SMPTE journal*, SMPTE, v. 109, n. 3, p. 178–187, 2000.

49 BORER, S.; LESZCUK, M. *MOAVI indicators*. Disponível em: <http://vq.kt.agh.edu.pl//metrics.html>.

50 LIU, L.; LIU, B.; HUANG, H.; BOVIK, A. C. No-reference image quality assessment based on spatial and spectral entropies. *Signal Processing: Image Communication*, v. 29, n. 8, p. 856 – 863, 2014. ISSN 0923-5965. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0923596514000927>.

51 MOORTHY, A. K.; BOVIK, A. C. A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*, v. 17, n. 5, p. 513–516, May 2010. ISSN 1070-9908.

52 AHMED, N.; NATARAJAN, T.; RAO, K. R. Discrete cosine transform. *IEEE Transactions on Computers*, C-23, n. 1, p. 90–93, Jan 1974. ISSN 0018-9340.

53 SCHULDT, C.; LAPTEV, I.; CAPUTO, B. Recognizing human actions: A local svm approach. In: IEEE. *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. [S.l.], 2004. v. 3, p. 32–36.

54 LIN, W.-H.; HAUPTMANN, A. News video classification using svm-based multimodal classifiers and combination strategies. In: ACM. *Proceedings of the tenth ACM international conference on Multimedia*. [S.l.], 2002. p. 323–326.

55 MICHEL, P.; KALIOUBY, R. E. Real time facial expression recognition in video using support vector machines. In: ACM. *Proceedings of the 5th international conference on Multimodal interfaces*. [S.l.], 2003. p. 258–264.

56 CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995.

57 HSU, C.-W.; LIN, C.-J. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, IEEE, v. 13, n. 2, p. 415–425, 2002.

58 KEYS, R. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 29, n. 6, p. 1153–1160, Dec 1981. ISSN 0096-3518.

59 GROUP, V. Q. E. et al. Final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii (fr_tv2). 2003.

60 PINSON, M. H.; WOLF, S. An objective method for combining multiple subjective data sets. *Proc. SPIE*, v. 5150, p. 583–592, 2003. Disponível em: <http://dx.doi.org/10.1117/12.509909>.

61 SHEIKH, H. R.; SABIR, M. F.; BOVIK, A. C. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, IEEE, v. 15, n. 11, p. 3440–3451, 2006.

62 GASTALDO, P.; REDI, J. A. Machine learning solutions for objective visual quality assessment. In: *6th international workshop on video processing and quality metrics for consumer electronics, VPQM*. [S.l.: s.n.], 2012. v. 12.

63 MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN 0521865719, 9780521865715.

64 SPEARMAN, C. The proof and measurement of association between two things. *The American journal of psychology*, JSTOR, v. 15, n. 1, p. 72–101, 1904.

APPENDIX

I. EVALUTUATION MEASURES

To evaluate classification methods is common to use the recall, precision an F1-Score [63]. The results of a classification method can be represented in a confusion matrix. The example of a confusion matrix in a binary classification is shown in Table I.1.

Table I.1: Example of a Confusion Matrix

	Real Positive	**Real Negative**
Predicted Positive	True Positive	False Positive
Predicted Negative	False Negative	True Negative

The confusion matrix represents all the results of a classifier into a single table. Predicted Positive are the cases, which the method classify as positive. Predicted Negative are the cases, which the method classify as negative. Real Positive are the cases that are labeled as positive. And Real Negative are the cases labeled as negative. The true positive are the cases which the classifier correctly predicted as positive or negative cases, respectively. The false positive and false negative are the cases which the classifier wrongly predicted as positive or negative cases, respectively.

I.1 RECALL

Recall measures the proportion of the real positive cases correctly classify by the classification method. It is most relevant in applications that aim to identify real positive cases. Recall is defined by:

$$Recall = \frac{\text{True Positive}}{\text{Real Positive}}, \tag{I.1}$$

I.2 PRECISON

Precision measures the proportion of the predicted positive cases correctly classify by the classification method. It is a measure of the true positive accuracy. Precision is define by:

$$Precision = \frac{\text{True Positive}}{\text{Predicted Positive}}, \tag{I.2}$$

I.3 F1-SCORE

F-Measure is a single measure that combines the results of precision and recall. F-Measure is a harmonic mean of precision and recall, define by:

$$F - Measure = \frac{1}{\alpha \frac{1}{Precision} + (1-\alpha)\frac{1}{Recall}} = \frac{(\beta^2+1)Precision \cdot Recall}{\beta^2 Precision + Recall} \tag{I.3}$$

where $\beta^2 = \frac{1-\alpha}{\alpha}$, β is a parameter that controls the balance of precision and recall. F1 Score is the special case when $\beta = 1$, in other words, the F-Measure equally weights precision and recall. F1 Score is defined by:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}, \tag{I.4}$$

II. SPEARMAN CORRELATION

The Spearman Correlation was proposed by Spearman in 1904 [64]. It measures the prediction monotonicity of a model. In this work, we use the Spearman Correlation to compare the prediction quality scores, given by the video quality assessment methods, with the subjective quality scores. Let x_i denote the predict score and y_i denote the subjective score, where $i \in 1, 2, ..., M$. The Spearman Correlation Coefficient SCC is defined by the following equation:

$$SCC = \frac{\sum_{i=1}^{M}(\chi_i - \bar{\chi})(\gamma_i - \bar{\gamma})}{\sqrt{\sum_{i=1}^{M}(\chi_i - \bar{\chi})^2}\sqrt{\sum_{i=1}^{M}(\gamma_i - \bar{\gamma})^2}}, \tag{II.1}$$

where χ_i is the rank of x_i and γ_i is the rank of y_i in the ordered data series, $\bar{\chi}$ and $\bar{\gamma}$ are the respect midrank. The SCC makes no assumption about the shape of the relationship between x_i and y_i.