Pontificia Universidad
JAVERIANA
Cali

FACULTY OF ENGINEERING

DEPARTMENT OF ELECTRONICS AND COMPUTER SCIENCES

Evaluation of no-reference quality prediction metrics in videos impaired by authentic distortions

José Alejandro Ledesma Mazuera
Stidl Alfonso Torres Morón

Supervised by
Hernán Darío Benítez Restrepo
Roger Alfonso Gómez Nieto

Santiago de Cali, JAN 13th, 2021

Abstract

Digital video processing and hardware systems can introduce distortions into the video signal during the capture process. Video quality assessment (VQA) is a key factor in the success of a multimedia system or service, which aims to make the quality of the experience perceived by the user acceptable. For this reason, in recent years has accelerated considerably the study and development of automatic objective methods that accurately quantify the impact of visual distortions in the perception without having as reference the original video. Verification of no-reference video quality algorithms requires realistic databases of distorted video and human judgments of these. However, most of the current publicly available video quality databases have been created under highly controlled conditions using simulated (artificial) and post-capture distortions in high-quality video. This situation motivates us to carry out this project, in which we evaluated state-of-the-art no-reference metrics such as FRIQUEE, QAWV, BRISQUE, NIQE, NSTSS, and TLVQM in authentically distorted video databases such as KoNViD-1k, LIVE-Qualcomm, and LIVE Video Quality Challenge (VQC). In addition, we evaluate a seventh VIIDEO algorithm and a fourth CVD2014 database. This project provides a systematic study on how the distortions in the capture are predicted by automatic models of perceptual quality and evaluated according to the correlation between these predictions and human quality evaluations. The repository with the algorithms and results is available at https://github.com/nomsedel/Evaluation-of-no-reference-qualityprediction-metrics-in-videos-impairedby-authentic-distortions.

***Index terms*—-** Video quality assessment (VQA), capture distortions, subjective quality assessment, objective no-reference algorithms, authentic distortions, databases, content characteristics, regression, temporal modeling, PLCC, SROCC and RMSE.

Acknowledgments

First of all, we are grateful to God for the good health and well-being, fundamental pillars to complete this degree work, to our parents for the incessant encouragement, support, and attention. We also wish to express our sincere gratitude to our supervisors, the Pontificia Universidad Javeriana Cali, for providing us with all the necessary tools for research.

We take this opportunity to express our gratitude to all the faculty members of the Department of Electronic Engineering for their help, support, and commitment to training excellent engineers. We also express our gratitude to each and every one of those who, directly or indirectly, have participated in our training as electronic engineers.

This work was supported by the MINCIENCIAS and Pontificia Universidad Javeriana Seccional Cali with the Project Vigilancia Inteligente para la red de cámaras de la Policía Metropolitana de Cali under Grant Project 1251-745-57892.

Contents

CHAPTER 1

Introduction

The use of video and multimedia applications is growing rapidly in everyday life. In the mass consumer market, different providers are offering video services and applications to end-users. In this scenario, it is essential to ensure an appropriate quality of experience (QoE) for the user, since every day, thousands of videos impaired by in-capture distortions are uploaded. The sources of these distortions are blurring, camera destabilization, and poor lighting [1]. Hence, it is necessary to predict the video quality (VQA) through subjective and/or objective studies. It is for this reason that the Laboratory for Image and Video Engineering (LIVE) at the University of Texas at Austin has carried out several subjective and objective studies taking into account the authentic distortions [2], among these, are the incorrect representation of colors, low exposure, lack of sharpness and overlapping of content in the video [1].

Video traffic already represents 80% of all the mobile internet traffic [10], these multimedia contents are distributed through the telecommunications networks experiencing various types of distortions or degradation's during the process of acquisition, compression, processing, transmission, and reproduction. These distortions are introduced by the camera hardware or processing software during the capture process, because video systems use focusing and compression techniques with loss of information, and the transmission media in turn can introduce distortion factors; such as delays, packet loss, among others [1].

That is why subjective methods are the most reliable way to measure the quality of an image or video, because it is done by a group of people who give their opinion about their perception, but these subjective studies are expensive, difficult to perform and impractical in real-time applications [11]. In recent years the design, the use and the study of automatic objective methods for video quality prediction have advanced, capable of reliably predicting the perceived quality, from objective measures taken at some point in the system [12]. However, most studies have been carried out with video quality databases created under highly controlled conditions, using simulated (not authentic) graded and post-capture distortions.

Therefore, this proyect raises the following research questions:

1. How do no-reference VQA metrics work in truly distorted databases?
2. Do no-reference VQA metrics correlate with the human ratings obtained in subjective studies?

In the chapter 1 of this document present the subjective and objective methods, authentic and in-capture video distortions, content-aware features, regression and data analysis measures. Chapter 2 describes the four databases to be analyzed in each of the seven NR VQA metrics presented in chapter 4, and in chapter 5 we will describe the process followed to obtain the results that were later analyzed in chapter 6, and finally we will present some conclusions in chapter 7 about the work done.

1.1 Objectives and Scope

1.1.1 General Objective

To predict the video quality of four VQA databases KoNViD-1k [30], LIVE-Qualcomm [1], LIVE Video Quality Challenge (VQC) [2] and CVD2014 [36], by applying seven no-reference state-of-the-art VQA metrics FRIQUEE [3], QAWV [4], BRISQUE [5], NIQE [6], NSTSS [7], TLVQM [8] and VIIDEO [43].

1.1.2 Specific Objectives

1. To extract video quality scores and features from four publicly available VQA datasets KoNViD-1k [15], LIVE-Qualcomm [1], LIVE Video Quality Challenge (VQC) [16] and CVD2014, according to seven no-reference video quality models.
2. To train a regressor for each no-reference video quality model to predict human scores of perceptual video quality.
3. To evaluate the performance of no-reference VQA metrics based on Pearson Linear Correlation Coefficient (PLCC), Spearman's Rank-Order Correlations Coefficients (SROCC), and Root Mean Square Error (RMSE).

The results of the objective study that compares seven no-reference metrics are particularly desirable in networked visual communication applications for the purpose of monitoring the quality of service (QoS) [9]. Image and video content delivered over various wired and wireless networks inevitably suffers degradation of visual quality during lossy compression and transmission over error-prone networks. It is imperative that network service providers monitor these quality degradations in real-time to optimize network resource allocations and maximize user expectations within certain cost constraints. It has been shown that typical error criteria used in network design and testing, such as binary error rate (BER), do not correlate well with the quality of the network consumer experience [9]. Therefore, accurate and high-speed measurements of VQA perception play an important role [9].

CHAPTER 2

Theoretical framework

2.1 State of the art

2.1.1 Subjective methods

The most reliable way to measure the quality of an image or video is through subjective evaluation, which is carried out by a group of observers who give their opinion about their perception of the video quality [21], [22]. In these subjective tests, video sequences are shown to a group of viewers. This viewer's opinion is recorded and averaged as the Mean Opinion Score (MOS), which is mathematically defined in the Equation 1.

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N}$$

Equation 1. Mean Opinion Score.

The MOS is calculated as the arithmetic mean of the individual scores given by people for a given stimulus in a subjective quality assessment test; where R_n is the individual scores given by the stimulus subjects and N is the total number of people who took the test. It should be noted that the observers are individuals who judge the quality based on their own perception and previous experience [13].

The quality scales used by the subjective methods can be continuous or discrete (typically between 5 and 11), depending on the case. In the case of MOS, the most widely used and accepted metric system is the 5-point scale (1 - 5), in which: 1 is the worst score, indicating "very poor" quality; 2 "poor"; 3 "fair"; 4 "good" and 5 "excellent" [14]. Subjective methods are expensive, difficult, and impractical to perform in real-time applications, so it is necessary to use and develop objective and automatic methods, capable of reliably predicting perceived quality.

2.1.2 Objective methods

Methods of objective VQA are mathematical and/or statistical models that approximate the results of subjective quality assessments. Objective methods are based on statistical criteria,

such as regressors or training metrics, which can be objectively measured and automatically evaluated by a computer program.

1. Pixel-Based Methods (NR-P) : pixel-based models use a coded representation of the signal and analyze quality based on the pixel information. Some of them evaluate only specific types of distortions, e.g. blurring or other coding artifacts.
2. Parametric/Bitstream Methods (NR-B) : these models use features extracted from the video bitstream, which can be packet headers, motion vectors and quantification parameters.
3. Hybrid Methods (Hybrid NR-P-B) : they are a mixture of the NR-P and NR-B models.

2.1.3 Authentic and In-capture Video Distortions

Videos captured with high-end cameras and then impaired by distortions introduced synthetically, (post-capture); compared to the previous one's videos captured by ordinary people are very different because they contain real-world distortions that are introduced by the many different mobile cameras on the market today, we refer to these latter videos as authentically distorted [16], [17].

The vast majority of mobile digital videos produced and consumed in social media are taken by casual, inexperienced users, and the capture process is often affected by sensitive variables such as lighting, exposure, lens limitations, noise sensitivity, acquisition speed, in-camera processing, and camera movement, each of which can adversely affect the perceived visual quality of a video. However, the latest generation of cameras often allow users to control some of the parameters of video acquisition, and the unsafe eyes and hands of most amateur camera users often result in the presence of annoying video distortions during capture, despite attempts to include corrective software in the camera devices [1].

A key aspect of the real world is that truly distorted video when captured by users who are inexperienced in using cameras; videos from such users cannot be accurately described as suffering from unique and separable distortions. Currently, there is no known way to categorize, characterize or model the complex and uncontrolled combinations of video distortions that occur in real-life scenarios [1].

2.1.4 Content-Aware Features

Content-aware features help address content dependency on the intended image and/or video quality to improve the performance of target models [18-21]. Initially, the relevant features of the videos to be studied are extracted in order to refine the existing quality measures [19], using the semantic information of the upper layer of the pre-trained image and/or video classification networks to incorporate them into the traditional quality features [20], [21]. This is done in order to exploit the aggregation of deep semantic features of multiple patches for quality assessment.

These deep semantic features have been shown to alleviate the impact of content on the quality assessment task [19], [4]. Inspired by this work [19], the features will be extracted from 2,227 videos (contained in the databases mentioned above) and then trained into a statistical regression model.

2.1.5 Regression

These features of the perceptually relevant videos, along with the corresponding real-value MOS of the training set, are used to train a support vector regressor. SVR is the most common tool for learning a non-linear mapping between the features of a frame and a single label (quality score) of objective quality assessment algorithms. Given a feature vector the Support Vector Machines (SVM) maps this high dimensional vector into a visual quality score [3].

SVR is widely used in many disciplines due to their high accuracy, their ability to handle high dimensional data, and flexibility in the modeling of various data sources [3]. Although VQA databases are available, they are not large enough to motivate the use of deep learning methods.

2.1.6 Data analysis measures

Linear Correlation of the Pearson Coefficient (PLCC), it is a linear measure between two quantitative random variables (x,y) [22]. Pearson's correlation is independent of the scale on which the variables are measured, and it is used to measure the degree of relationship between two variables, provided that they are quantitative and continuous, in which case it is necessary to calculate the PLCC of the indices of subjective quality with the objectives

[23]. Equation 2 is used to calculate the PLCC for a population.

$$\rho_{xy} = \frac{E[(x - \mu_x)(y - \mu_y)]}{\sigma_x \sigma_y}$$

Equation 2. PLCC on a population.

With σ_x is the standard deviation of the variable x, σ_y is the standard deviation of the variable y, μ_x and μ_y are the mean of the variable x and y, and E is the expected value [22]. On the other hand, Equation 3 is used to calculate the PLCC for a statistical sample

$$rxy = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

Equation 3. PLCC on a sample statistic.

It is known that the value of the correlation index varies in the interval [-1.1] so if it is obtained [22]:

1. A PLCC equal to 1 means that when one of the variables increases, the other also increases in a constant proportion, while a PLCC equal to -1 means that when one of the variables increases, the other decreases in a constant proportion.
2. A PLCC between 0 and 1, there is a positive correlation; and if it is between -1 and 0, there is a negative correlation.
3. If the PLCC is equal to 0, there is no linear relation.

As a hypothesis the PLCC is expected to be as close to 1 as possible, because the objective NR quality assessment methods make a good prediction of the video quality, so their regressors and other statistical models work well.

Spearman's Rank-Order Correlation Coefficient (SROCC), is a measure of correlation or interdependence between two random variables X and Y, such variables can be both continuous and discrete [24]. SROCC is denoted by the symbol rho (or the Greek letter p, pronounced rho) and it is calculated as Equation 4.

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

Equation 4. Spearman's Rank-Order Correlation Coefficient (SROCC).

To calculate ρ, data are sorted and replaced by their respective order. D is the difference between the corresponding $X - Y$ order statistics. N is the number of data pairs [26].

It should be noted that the existence of identical data should be taken into account when ordering them, although if they are few, this can be ignored [26]. Finally, the interpretation of SROCC is the same as that of PLCC.

Square Root of the Mean Square Error (RMSE), it is the square root of the variance, i.e. the standard deviation. Its function is to measure the mean of the square errors, in other words, the difference between the estimator and what is estimated [39]. It is calculated as Equation 5:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Xi - Yi)^2}$$

Equation 5. Root Mean Square Error (RMSE).

In which X_i is the vector of the objective quality scores given by the NR metrics, Y_i are the subjective scores and n are the number of elements of vector x or y, because both must have the same cardinality.

CHAPTER 3

Datasets

The databases, composed by videos authentically distorted, and used for testing the NR VQA algorithms are: KoNViD-1k [15], LIVE-Qualcomm [1], LIVE Video Quality Challenge (VQC) [16] and CVD2014 [35] typically used in these studies.

3.1 KoNViD-1k

It consists of a VQA database with 1200 public domain video sequences, with videos encoded at three frame rates: 24, 25 and 30 Fps corresponding to 27%, 5% and 68% of the videos in the database, respectively [30]. In addition, it has a total of 12 resolutions, where the highest percentage (85%) of the videos have a frame size of 1280x720 pixels, followed by 1920x1080 (9%), and most of the videos (97%) have an audio channel [30]. Due to the large number of videos, the subjective scores were obtained using the CrowdFlower platform in crowdflow [30].

To build the KoNViD-1k database, it tooks as reference the YFCC100m database, which contains 793.436 video sequences from Creative Commons (CC) [30], which were filtered taking first into account the following aspects [30]:

- They were still available for download.
- They played at more than 15 frames per second (FPS).
- Lasted longer than 8 seconds.
- The videos did not have a "No Derivative Works" CC attribute.
- They had a resolution higher than 960×540 (W×H).
- They were in landscape layout.

After this filtering, a selection was made depending on 6 attributes related to the temporal, color, and spatial aspects in order to measure the diversity of content in the video databases

[30].

1. Blur: The blur of a frame was evaluated using the cumulative probability of blur detection (CPBD) metric [31].
2. Colorfulness: Hasler and the Suesstrunk metric reported in [32] were used for this attribute.
3. Contrast: The contrast of the picture was measured simply by the standard deviation of the pixel grayscale intensities [33]. The average of the standard deviation of the frame level then gave the contrast of a video.
4. Spatial information: The spatial information (SI) was obtained by applying a Sobel filter to each frame to extract the magnitude of the gradient for each pixel and then calculating its standard deviation [34]. The average standard deviation of all the frames resulted in the SI of the video.
5. Time information: Similar to SI, time information (TI) is the average of the standard deviations of the frame difference per pixel [34].
6. VNIQE: The Natural Image Quality Evaluator (NIQE) [35] was used as a substitute to evaluate the quality of the video by calculating the average NIQE value of all the frames.

The filtering thresholds for each attribute, as shown in Table 1, were chosen empirically on the basis of a qualitative inspection, so that most of the filtered videos show obvious artificial content (videos representing unnatural scenes such as screenshots or stop-motion sequences, as well as videos that are too dark or too bright) [30].

Attribute	**Lowest value**	**Highest value**
Blur amount	0.005	0.88
Colorfulness	4.37	123
Contrast	7.51	97.48
Spatial information	7.7	187.76
Temporal information	3.07	56.81
VNIQE	3.58	23.08

Table 1. Attribute thresholds for filtering outlier videos [30].

Figure 1 below shows 4 videos extracted from the KoNViD-1k database, where it can be seen that low scores are usually caused by strong motion blurring due to jolts or blurs and high score videos are sharp and show little distortion [30].

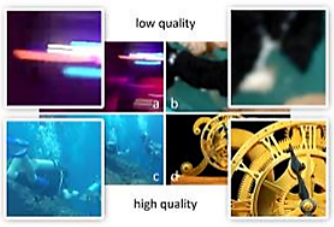

Figure 1. Example extreme quality videos: a. MOS 1.26 (lowest quality score in KoNViD-1k), b. MOS 1.52, c. MOS 4.12, d. 4.64 (highest score) [30].

3.2 LIVE-Qualcomm

This database has a total of 208 videos, in which all the videos were captured at a resolution of 1920x1080, and were also captured in environments where the acquired videos were affected by any of the following six distortions: Artifacts, color, exposure, focus, sharpness, and stabilization as shown in Figure 2 [1].

- Artifacts: Noise and blockiness distortions not part of the video content [1].
- Color: Videos with incorrect or insufficient color representation.
- Exposure: Over/under-exposure, making it difficult to see parts or the entirety of the scene [1].
- Focus: Auto focus related distortions, i.e., videos that are intermittently sharp or blurry over time [1].
- Sharpness: General unsharpness, i.e., lack of detail, texture, or sharpness. This distortion differs from out-of focus distortion in that with sharpness distortion, objects are in focus but do not appear 'crisp' or detailed [1].
- Stabilization: Camera shake that overwhelms content [1].

Figure 2. Sample frames of the unique video contents contained in LIVE-Qualcomm. Nine unique contents were captured per distortion category (illustrated along each column) [1].

The videos in this database were captured from the following eight mobile devices (as shown in Table 2): Samsung Galaxy S5, Samsung Galaxy S6, HTC One VX, Apple iPhone 5S, Nokia Lumia 1020, LG G2, Samsung Galaxy Note4 and Oppo Find 7, limited to take almost identical natural scenes because the field of view (FoV) is different in each device [1].

Phone	Artifacts	Color	Exposure	Focus	Sharpness	Stabilization	Total number of videos
Samsung Galaxy GS5	2	3	3	4	6	4	22
Samsung Galaxy GS6	8	6	5	6	3	5	33
HTC One VX	8	6	5	6	2	5	32
Apple iPhone5s	1	3	4	3	6	4	21
LG G2	7	6	5	5	3	5	31
Nokia Lumina 1020	2	2	3	2	4	3	16
Samsung Galaxy Note4	1	3	4	3	7	4	22
Oppo Find 7	7	6	5	5	3	5	31
Total	36	35	34	34	34	35	208

Table 2. Number of videos per phone and distortion. The rows are indexed by the eight mobile phones used to capture video content. The columns are indexed by the six dominant distortion categories [1].

3.3 LIVE-VQC

LIVE Video Quality Challenge (VQC) contains 585 unique content videos, captured on 101 different devices (43 device models) by 80 users with wide ranges of complex and authentic

distortion levels. These users uploaded their videos as they were captured, without any processing (for example, using video processing "applications" such as Instagram or Snapchat). Only videos with a duration of at least 10 seconds were accepted. No instructions were given on the content or style of capture [2].

The content of the videos is quite diverse, and includes scenes from sports games, music concerts, nature, various human activities (parades, dancers, street artists, cowboys, etc.), and much more, as shown in Figure 3. The scenes were captured under different lighting conditions (different times of day and night), and include both indoor and outdoor scenes. Very diverse levels of movement (camera movement and frame movement) are present and often contribute to complex, space-distorting variations. It contains a large number of subjective video quality scores through the crowdsourcing tool, with an average of 240 human opinions recorded per video [2].

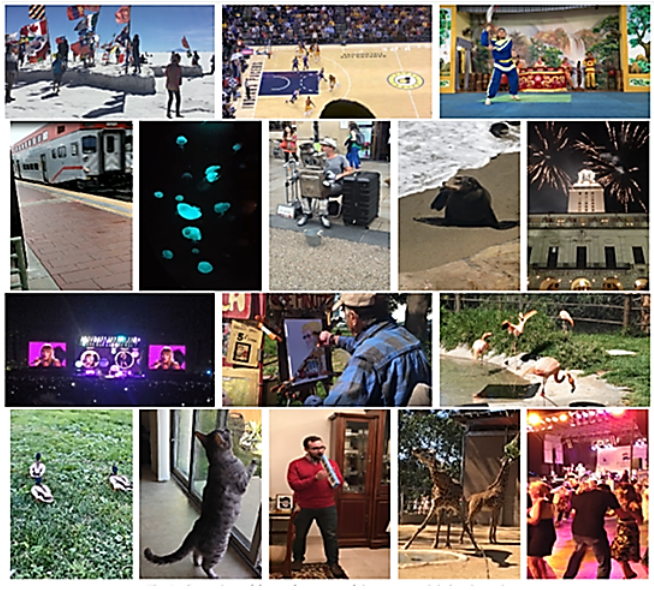

Figure 3. Screenshots of frames from some of those presented during the study [2].

A taxonomy of mobile devices used to capture video is given in Table 3. There were 101 different devices deployed (some users provided videos captured by multiple devices), including 43 different models. Commercial launches of the devices varied between 2009 and 2017, although most of the videos were captured by devices launched after 2015 and beyond [2].

Make	Model	Number of videos
Amazon	Fire HDX	1
Apple	Ipad Pro	2
Apple	Iphone 3GS	1
Apple	Iphone 4	14
Apple	Iphone 4s	2
Apple	Iphone 5	25
Apple	Iphone 5s	49
Apple	Iphone 6	48
Apple	Iphone 6s	107
Apple	Iphone 6s plus	5
Apple	Iphone 7	17
Apple	Iphone 7 plus	13
Apple	Ipod touch	8
Asus	Zenfone Max	1
Google	Pixel	7
Google	Pixel XL	20
Hisense	s1	1
HTC	10	13
HTC	M8	5
Huawei	Nexus 6P	10
LG	G3	3
LG	G4	2
LG	Nexus 5	50
Motorola	E4	1
Motorola	Moto G 4G	3
Motorola	Moto G4+	1
Motorola	Moto Z Force	12
Nokia	Lumina 635	5
Nokia	Lumina 720	3
OnePlus	2	4
OnePlus	3	4
Samsung	Core Prime	5
Samsung	Galaxy Mega	1
Samsung	Galaxy Note 2	21
Samsung	Galaxy Note 3	5
Samsung	Galaxy Note 5	72
Samsung	Galaxy S3	4
Samsung	Galaxy S5	25
Samsung	Galaxy S6	14
Samsung	Galaxy S8	6
Xiaomi	MI3	1
Xperia	3 Compact	3
ZTE	Axon 7	1

Table 3. Number of videos captured by each type of camera devices [2].

No restrictions were placed on the orientation of the camera device during or post-capture, and 23,2% of the videos in the database were taken in portrait mode and the other 76,2% in landscape mode [2]. On the other hand, the predominant resolutions in this database are

1920×1080, 1280×720 and 404×720, which together represent 93,2% of the total, as shown in Figure 4 [2].

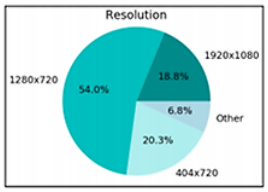

Figure 4. Distribution of the resolutions of the LIVE Video Quality Challenge database [2].

3.4 CVD2014

CVD2014 contains 234 videos of five different scenes captured by 78 different cameras (cell phones, compact camera, video camera, SLR), in which the subjective experiments were performed following the single stimulus procedure (SS) to collect the video quality classifications [36].

The video sequences in the CVD2014 database were captured from 5 different scenes. Figure 5 shows three frames of the scenes. The frames are from the beginning, middle and end of the video sequences. The length of the clipping and processed videos was from 10 to 25 s [36].

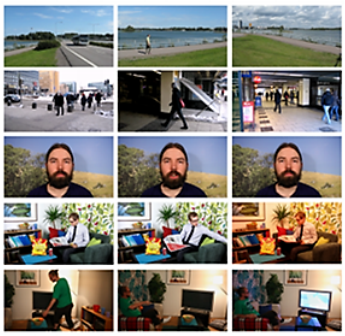

Figure 5. CVD2014 video database sequences 1-5 (from top to bottom): Traffic (1), City (2), Talking Head (3), Newspaper (4) and Television (5) [36].

The 5 scenes are described below:

1. Traffic: A bus is driving on a busy road and passes the camera. The camera pans to the direction of the sea where a man is walking on a walkway [36].

2. City: A view from a central location in a city where a man is walking from the outdoors to a tunnel, which includes a gradual change in color temperature and illuminance based on the panning camera and moving objects [36].
3. Talking Head: The upper body of a man who is talking (in Finnish) [36].
4. Newspaper: A man is reading a newspaper indoors, and the light turns to a different color temperature [36].
5. Television: A man is walking to a sofa and picks up an orange from a basket, sits down and switches on a TV, on which a news program begins [36].

These scenes included sharpness, graininess, color balance, darkness and jerkiness, as shown in Figure 6 [36].

Figure 6. Example frames from typical video sequences in the CVD2014 database. Descriptions of the video sequences given by subjective observers: (a) sharp and bright, (b) grainy, (c) unsharp and yellow, (d) sharp and bright, (d) unsharp and reddish, and (e) unsharp, shivery, grainy and dark [36].

Since the scenes contain different amounts of spatial and temporal information, the metrics of spatial perceptual information (SI) and temporal perceptual information (TI) are used to measure the level of activity in a video sequence as shown in Figure 7 [36].

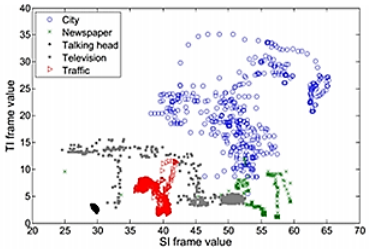

Figure 7. Spatial and temporal activity presented as point cloud values for the example (high-quality) video sequences[36].

The videos were calculated in 10 different formats because 78 different devices were used. Table 4 shows the video formats used according to the frames per second [36].

Video format	**10**$\angle fps \leq$ 13	**13**$\angle fps \leq$ 16	**19≤fps≤22**	**22≤fps≤25**	**25≤fps≤28**	**28≤fps≤31**
QCIF (176 x 144)		1				
QVGA (320 x 240	1					1
CIF (352 x 288)		1				
VGA (540 x 480)		7			1	18
NTSC (720 x 480)			1	1		2
PAL (768 x 576)				3		
WPAL (848 x 480)				1		
HD (1080 x 720)				8	1	9
FHD (1920 X 1080)				6	1	15

Table 4. Frequency table of video formats and frame rates in the cameras used to capture the videos in the CVD2014 database [36].

Table 5. Public consumer video quality databases compared: KoNViD-1k, LIVE-Qualcomm, LIVE-VQC [14] and CVD2014 [36].

Dataset characteristic	KoNViD-1k [15]	LIVE-Qualcomm [1]	LIVE-VQC [16]	CVD2014 [36]
Number of test videos	1200	208	585	234
Video resolution	960x540	1920x1080	320x240-1920x1080	640x480, 1280x720
Video frame rate	23-29 frames/sec	30 frames/sec	19-30frames/sec	9-30 frames/sec
Video length	8 sec	15 sec	10 sec	11-28 sec
Number of scenes	1200	54	585	5
Number of devices	164	8	101	78
Test methodology	Crowdsourcing	Lab-based	Crowdsourcing	Lab-based
Number of test subjects	642 (min. 50 per video)	39	min. 200 per video	27-33 (six experiments)
Rating scale	Absolute Category Rating (1-5)	Continuous 0-100	Continuous 0-100	Continuous 0-100
Audio track included	Yes (some)	No	Yes	Yes
Main strength	Very wide diversity of contents and distortion types. Large number of test users.	Realistic consumer content with smartphones. Uses Full HD resolution.	Realistic consumer content with a wide diversity of scenes. Large number of test users.	Realistic consumer content with a large number of devices and impairment types.
Main weakness	Some contests and distortions have little practical relevance to NR-VQA. Test methodology prone to unreliable individual scores.	Large number of scenes, but different scene types not very well balanced. Only smartphones used as camera.	Distribution of MOS values biased towards high scores. Some resolutions represented by few sequences only.	Some methodological inconsistencies between experiments. Small number of different scenes.
Remarks	Some contents in the database are clipped from the original.	Additional information collected concerning the dominating distortion type of each test sequence.		Six different experiments. In some experiments, other information collected aside of video quality (audio quality, contrast, blurriness etc.)

CHAPTER 4

Video quality assessment (VQA) metrics

Algorithms can be created that learn the human responses to distortion by training them in large databases of human opinion scores [37], which is why for this project we used seven NR image quality assessment (IQA) and VQA metrics:

4.1 Quality Assessment of In-the-Wild Videos

Quality Assessment of In-the-Wild Videos (QAWV) is a No-Reference method of predicting video quality that takes into account the effects of content dependence as shown in Figure 8 and temporal memory (long-term dependence and temporal hysteresis). This is because humans tend to remember poor quality frames in the past and thus decrease the perception of quality for subsequent frames, even when the quality of the frame has returned to acceptable levels [4,38].

The framework of the method is shown in Figure 8. The fist stage extracts the characteristics of each video frame through a CNN (convolutional neural network) called Resnet-50 which was pre-trained on imageNet [39], it yields a feature map with a dimension of 4096, and then apply a global pooling (GP) that consists of a mean and a global standard deviation grouping of the characteristics, the latter in order to preserve the information of the variation [4].

These characteristics are then concatenated and sent to a fully connected layer (FC) to perform the dimensional reduction to 128 in order to optimize the training of the GRU network (recurrent neural network), this is responsible for reducing dependencies in the long term and produce the quality of the frames. Finally, to take into account the temporal hysteresis effect, the overall quality of the video is grouped from these frames by a temporal reserve layer inspired by subjectivity [4].

Specifically, in this last stage a memory quality element is defined as the minimum of the quality scores in the previous tables. A current quality element is defined according to the ranking based on the weighted average order of the quality scores in the following tables; the estimated quality scores are calculated as the weighted average of the current memory and elements; the video quality is calculated as the temporary average of the approximate scores [4].

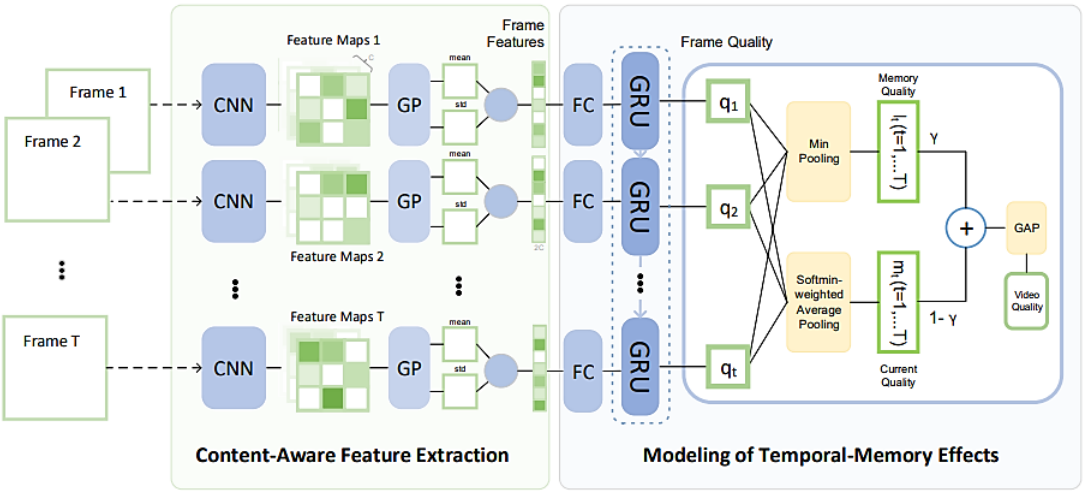

Figure 8. The overall framework of the proposed method [4].

4.2 TLVQM

Two-Level Video Quality Model (TLVQM) is an no-referenced method of predicting video quality, focused on extracting features intuitively related to the content of each video frame and its motion characteristics typically considered to be disturbing. This is because the perception of video quality is based on the interaction of temporal and spatial characteristics of the video, and camera movement is often the dominant type of distortion in the captured video [8].

It should be noted that the HVS is able to compensate for the blurriness of moving objects [40], since motion-induced blurriness masks spatial distortions [8].This model first divides the video sequence into segments, ie, if the video frame rate is 30 frames per second, a segment will contain 30 frames, as shown in Figure 9 (there are ten frames per segment) [8].

Next step, the vector of scalar values LCF (low complexity characteristics per frame) is calculated, using a grouping of three frames so that one frame is the reference frame (the square colored frame in Figure 9) and neighboring frames are the target frames for motion estimation (the light colored frames in Figure 9) [8].

Then, we proceed to calculate the time grouping by calculating the averages and standard deviations of each FLC characteristic; this grouping is the consistency of the segment level [8]. The frame with the most similar frame-level FLC vector compared to the mid-segment level FLC vector is selected as a representative frame of the segment and is used to calculate the spatial FHC [8], this is because the spatial characteristics of the frames tend to be highly correlated, and therefore we can assume that the FCH calculated from the representative frame gives a reasonable approximation of the general characteristics of spatial quality of the entire segment [8].

Finally, the mean temporal clustering is applied to the FCH, HCF and consistency characteristics of the segment level to obtain the characteristic vector of the entire video sequence, as shown in Figure 10 [8], and so these characteristics are used to train a regressor to obtain a prediction of objective video quality [8].

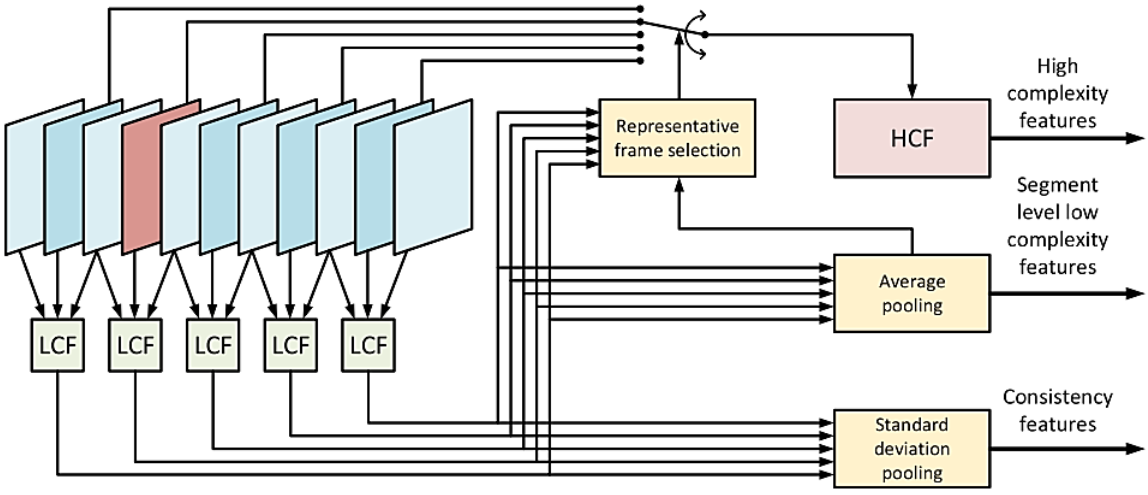

Figure 9. Processing of one segment of video to generate all types of features [8].

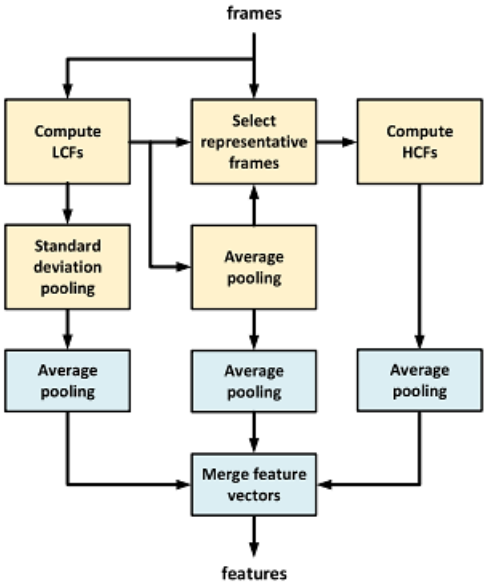

Figure 10. Block diagram of the proposed method for the sequence level feature vector computation [8].

4.3 NIQE

The Natural Image Quality Evaluator (NIQE) is an IQA NR OU (opinion unaware) - DU (distortion unaware) model, based on the extraction of "quality-aware" characteristics, which conform to a multivariate Gaussian model (MVG) [6].

First, as shown in Figure 11, the NIQE has a pristine image data store, where each of these images is divided into blocks and the NSS is calculated for each block [7]. The classic spatial model of the NSS [10] starts with the preprocessing of the image by removing the local mean and dividing normalization processes, to include only the blocks with statistically significant characteristics [6]. After this, the multivariate Gaussian mean and the standard deviation derived from the SSS characteristics are stored [41].

To calculate an image quality score for an arbitrarily distorted image, the SNS entities are extracted from the statistically significant blocks of the distorted image, and then adjusted to a multivariate Gaussian distribution to the SNS entities of the image [6], Therefore, the quality of a given image is then expressed as the distance between a multivariate Gaussian

adjustment (MVG) of the SSS characteristics extracted from the image, and an MVG model of the conscious quality characteristics extracted from the natural image corpus [6].

It should be noted that the NIQE index is not linked to any specific distortion, however, it offers an almost comparable predictive power for the same distortions in which the BRISQUE index has been trained, with a similar low complexity [6]. Finally, this model does not work well when images do not come from a natural source (e.g. computer graphics) or when natural images are subject to unnatural distortions [6].

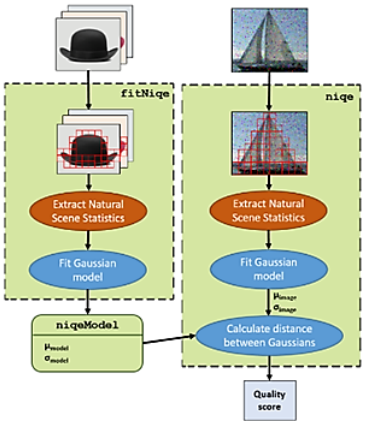

Figure 11. NIQE Workflow Diagram [41].

4.4 BRISQUE

Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) is a blind/NR-reference IQA evaluation metric that operates in the spatial domain and uses the scene statistics of locally normalized luminance coefficients to quantify possible naturalness losses in the image due to the presence of distortions, resulting in a holistic measure of quality [5]. BRISQUE uses an RVS trained by a set of characteristics derived from the empirical distribution of luminance and locally normalized luminance products, taking into account a spatial statistical model of the natural scene [5].

BRISQUE is limited to measuring the quality of images with the same type of distortion as those stored, since it has a data warehouse containing images with known distortions and pristine copies of those images, in addition to having a subjective opinion score for each

distorted image stored [41].

This metric calculates the NSS entities for each image, without dividing the image into blocks, as shown in Figure 12. These characteristics are derived from the empirical distribution of luminances and locally normalized luminance products, where the function uses these characteristics and the corresponding opinion scores to train a supporting vector machine regression model. In which the returned model stores the parameters of the support vector regressor [41].

BRISQUE is trained using subjective opinion scores so that the BRISQUE score correlates well with human perception of quality [41]. To calculate the quality score of an image with the same type of distortions, the NSS entities of the distorted image are extracted and a quality score is predicted by the support vector regression [41].

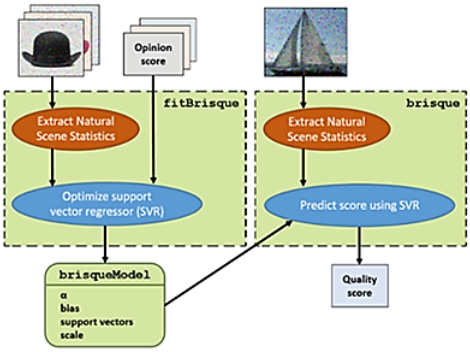

Figure 12. BRISQUE workflow diagram [41].

4.5 FRIQUEE

Feature Maps-Based Referenceless Image Quality Evaluation Engine (FRIQUEE) is a blind IQA model that proposes a feature mapping approach that avoids determining the type of distortion contained in an image, so FRIQUEE focuses on capturing the consistency or deviation of distortions. To do this, it combines a large and diverse collection of statistical features of perceptually relevant real-world images through 4 layers [3].

These layers are called FRIQUEE-Luma, FRIQUEE-Chroma, FRIQUEE-LMS and FRIQUEEALL. These layers extract feature maps focused on the luminance components (grayscale version) such as the luminance map, the neighboring paired products, the sigma map, the Gaussian difference (DoG) of the sigma map, the laplacian of the luminance map, the yellow channel

map and the features extracted in the wavelet domain, as shown in the Figure 13 [3].

These layers also make use of the color space of the LMS which is the most natural perceptual color space as it mimics the responses of the three types of cones in the human retina. The LMS space refers to the wavelength ranges to which the different cone types are sensitive: Long, Medium and Short. Knowing the response curves of these receptors, colors can be represented directly by making linear combinations of the components of the LMS [3].

The last layer called FRIQUEE-ALL uses all of the above feature maps, as well as the HSI color space feature map (Hue, Saturation, Lightness) and the yellow channel map [3]. After capturing these feature layers, a divisive normalization process is performed [3].

These features of each layer are given as input to the neural network and are used with the Support Vector Regression (SVR) for image quality prediction, as shown in the Figure 14 (IQA model). In addition, FRIQUEE is based on a multivariate Gaussian distribution and contains 330 statistical features of the natural scene that capture a richer set of true image distortions [3]. On the other hand, FRIQUEE is a model based on natural scene statistics (NSS) that is based on the hypothesis that the different existing statistical image models capture distinctive aspects of the loss of perceived quality of a given image [3].

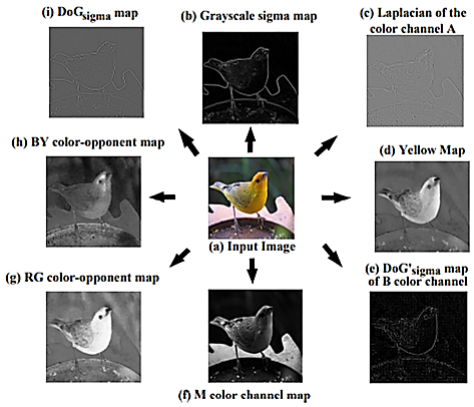

Figure 13. Given any image (a) FRIQUEE first constructs several feature maps in multiple color spaces and transform domains (some are shown here), then derives and extracts scene statistics from these maps after performing perceptually significant divisive normalization [42] on them.

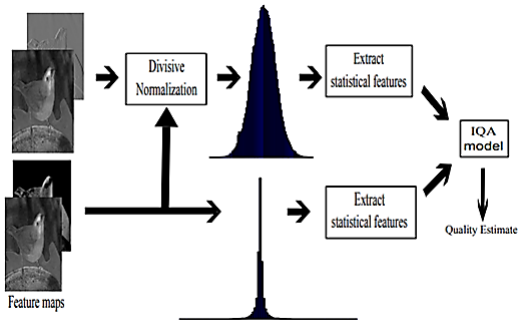

Figure 14. The proposed FRIQUEE model processes all the feature maps by modeling the distribution of their coefficients using either one of GGD (in real or complex domain), AGGD, or wrapped Cauchy distribution, and extracting perceptually relevant statistical features that are used by a quality predictor [3].

4.6 NSTSS

No-Reference Video Quality Assessment Using Natural Spatiotemporal Scene Statistics (NSTSS), is a VQA NR metric that proposes a video representation based on a parameterized statistical model for the spatial-temporal statistics of the mean and normalized contrast subtraction coefficients (MSCN) of natural videos [7]. Specifically, a generalized asymmetric Gaussian distribution (AGGD) is proposed to model the statistics of the MSCN coefficients of natural videos and their filtered bandpass outputs of spatial and temporal Gabor [7].

The parameters of the AGGD model serve as good representative characteristics for distortion discrimination. therefore, a supervised learning approach using support vector regression (SVR) is proposed in order to address the problem of no-reference video quality assessment (NR VQA) [7], offering a competitive performance in traditional (synthetic) distortions and an acceptable performance in true distortions [7].

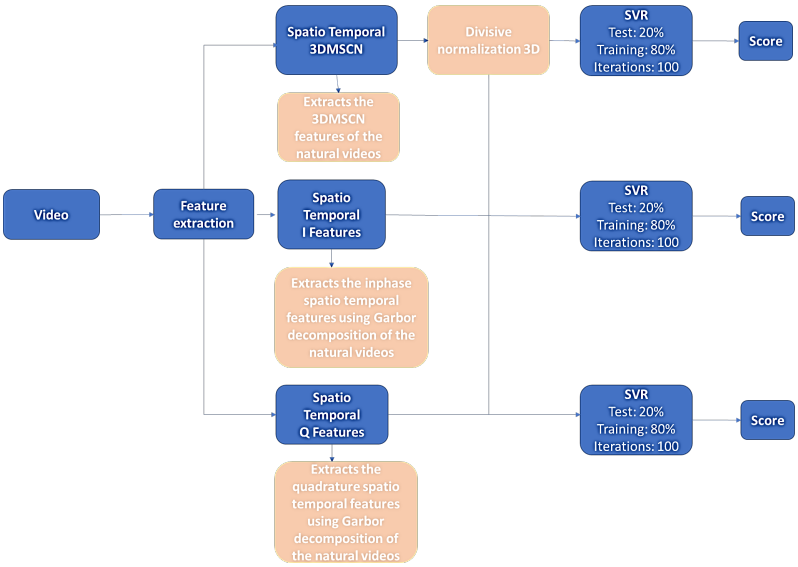

Figure 15. NSTSS workflow diagram.

4.7 VIIDEO

The video intrinsic integrity and distortion evaluation oracle (VIIDEO), is a VQA NR evaluation metric, capable of predicting the quality of distorted videos without any external knowledge about the pristine source, anticipated distortions or human judgments about the quality of the video [43]. This metric is based on a set of statistical models of perceptibly relevant temporal video difference signals from video frames [43].

The model calculates bandpass filter coefficients of video frame differences that capture temporary statistical regularities arising from structures such as moving edges [43]. When the coefficients have been calculated from good quality naturalistic still images (instead of frame differences), it has been observed that the coefficients reliably follow a generalized Gaussian distribution law [43].

On the other hand, when in natural video frames the differences are subjected to the common and unnatural distortions, their processed coefficients no longer tend to Gaussianity [43]. Since motion-induced changes are associated with strong correlations between the frame difference coefficients transformed into fine and coarse scales over time, while time-distortion-induced changes tend to cause transient statistical aberrations. VIIDEO captures

these differences by measuring the temporal variations of the correlations of the characteristics of the statistical model among the different scales of the coefficients of the locally transformed image [43]. Figure 16 summarizes the above using a workflow diagram.

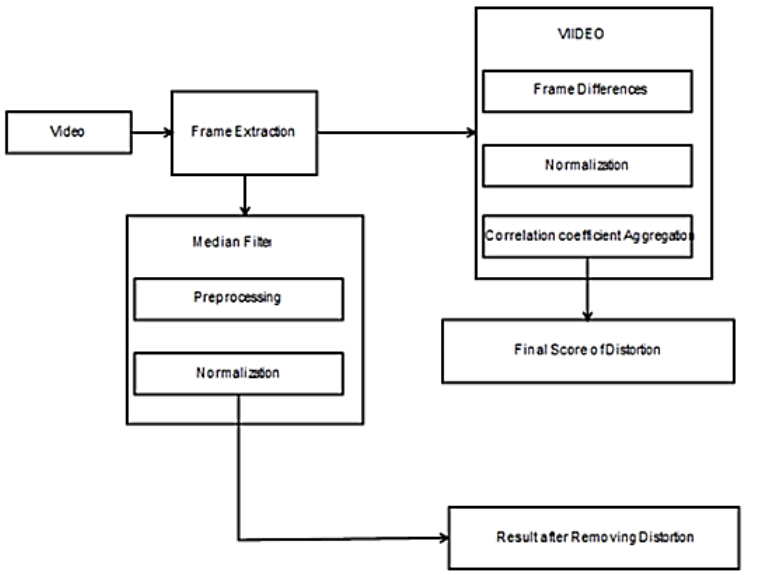

Figure 16. VIIDEO workflow diagram [44].

CHAPTER 5

Code optimization and execution times

5.1 FRIQUEE Optimization

For the FRIQUEE metric, tests were carried out with which it was estimated that the processing of each video would be around 6 hours, now considering the number of videos to be evaluated 2.227 is obtained that the processing time is very high so it was decided to reduce the calculation time of this algorithm, for this purpose tests were carried out on a Lenovo G40 I5 calculating the execution time for the metric. An image with a resolution of 500x500 pixels was introduced as an input parameter.

After this analysis it was found that the function "MGGD-ParaEstimate f" is the one with the longest computation time because it performs arithmetic operations between vectors, calling it 145 times, therefore, the operations were simplified, reducing the number of calculations, making use of variables in equations where they were repeated. After this, the function "DIIVINE-feature WrapCauchyEstimate" is the second with more individual execution time, receiving in total 96 calls, the number of while iterations was reduced because the results do not tend to vary if the iterations are higher than 10. Then in the main function "example", it was taken out the command "load" that loads the for learning model, so that it does not load it in all the iterations. Finally when making these changes it was managed to reduce the time of calculation in 26,514%.

5.2 Selection of development versions of BRISQUE and NIQE

For the BRISQUE and NIQE metrics there are versions developed under the Python and Matlab compilers. In order to use the optimal version in terms of execution time, performance and accuracy, 10 pristine images with different resolutions were selected, and execution times were calculated, as well as objective scores, as indicated in Table 6 and Table 7.

		Matlab		Python code		Python function	
Image	**Resolution**	**Score**	**Running time (s)**	**Score**	**Running time (s)**	**Score**	**Running time (s)**
A (1)	800 x 600	11.9127	0.5932	9.4164	0.239	9.2023	3.4031
A (2)	3000 x 4000	3.8309	5.0255	13.3787	5.0494	11.0781	101.0615
A (3)	768 x 512	55.6424	0.4131	64.7451	0.1559	68.2105	2.3657
A (4)	1024 x 683	28.7367	0.5651	30.3331	0.3064	28.0569	5.0571
A (5)	500 x 500	17.3495	0.3565	31.0315	0.1574	30.7983	1.9645
A (6)	500 x 500	26.3463	0.4821	27.737	0.1487	26.1521	1.8971
A (7)	768 x 512	97.4301	0.4816	99.5159	0.2738	28.1648	1.897
A (8)	768 x 512	32.0435	0.4734	37.6013	0.1878	35.4873	2.4681
A (9)	768 x 512	68.9316	0.5595	73.0834	0.2438	78.9015	3.5139
A (10)	768 x 512	43.1452	0.4821	45.5843	0.1917	47.046	2.7578

Table 6. Analysis of run times and scores in the Python and Matlab compilers of the BRISQUE metrics

		Matlab		Python	
Image	**Resolution**	**Score**	**Running time (s)**	**Score**	**Running time (s)**
A (1)	800 x 600	2.959	0.512	11.361	0.24
A (2)	3000 x 4000	4.586	12.176	12.059	5.795
A (3)	768 x 512	5.971	0.364	14.838	0.2
A (4)	1024 x 683	2.657	0.624	13.356	0.361
A (5)	500 x 500	5.663	0.217	27.562	0.163
A (6)	500 x 500	5.086	0.263	13.961	0.118
A (7)	768 x 512	55.558	0.358	42.059	0.235
A (8)	768 x 512	8.036	0.319	14.271	0.223
A (9)	768 x 512	17.865	0.332	23.904	0.216
A (10)	768 x 512	5.029	0.86	19.351	0.157

Table 7. Runtime and score analysis in the Python and Matlab compilers of the NIQE metric.

Since the execution times in the Matlab compiler are lower than those of the Python compiler, it was decided to work with this compiler, in addition to the fact that this was the one used in [5][6].

5.3 Execution times for VQA NR metrics

In order to evaluate the performance of the metrics regarding the resolution and number of frames of the videos with true distortions, four videos were selected, one for each database,

as indicated in Table 8 and Table 9.

	Running time of VQA Method (minutes)					
Videos	**BRISQUE**	**FRIQUEE**	**TLVQM**	**QAWV**	**NIQE**	**VIIDEO**
CVD-Test02-City-D01.avi	2.1804	208.5647	1.7054	5.2901	1.8190	5.5554
KoNViD-1k-5927908371	1.3392	165.2924	1.4136	3.9613	1.5199	3.4857
LIVEVQC-ABP-002-A069	4.3555	965.8803	6.8666	23.7258	8.3332	16.9229
Focus-Qualcomm-1006-ComplexTrain-HTCOneVX-VIDEO0087	6.8576	1368.8220	12.7618	34.7808	11.5214	23.0037

Table 8. VQA metrics execution times. according to 4 videos selected from each database (1 out of 1).

Videos	**Number of frames**	**Resolution**	**Duration of video (seconds)**	**Data rate (kbps)**
CVD-Test02-City-D01.avi	460	640x480	15	14747
KoNViD-1k-5927908371	240	960x540	8	9670
LIVEVQC-ABP-002-A069	300	1920x1080	10	9519
Focus-Qualcomm-1006-ComplexTrain-HTCOneVX-VIDEO0087	443	1920x1080	14	1493034

Table 9. Characteristics of the 4 videos selected from each database(1 out of 1) to calculate their execution times in the VQA metrics.

It should be noted that the execution times were calculated with an Acer E5-575G-56GB computer with the following hardware modifications: 500 Gb solid disk and 16 Gb RAM.

5.4 Procedure

5.4.1 Extraction and management of objective scores

Once the task of optimization of the metrics was completed, we proceeded to extract the characteristics and objective scores of each of the algorithms proposed in the objectives and scope of this project which are FRIQUEE [3], QAWV [4], BRISQUE [5], NIQE [6], NSTSS [7] and TLVQM [8] in the databases KoNViD-1k [30], LIVE-Qualcomm [1] and LIVE Video Quality Challenge (VQC) [2]. In addition, in order to go a little further than presented in the proposal of this project, it was decided to evaluate a seventh VIIDEO algorithm [43] and a fourth CVD2014 database [36].

The next step was to make some modifications to the algorithms to convert them from IQA to VQA, always keeping in mind not to alter the final result. Once the information was extracted from each video, it was stored in a Mat file for later use.

The sizes of the feature maps (used as arguments in the SVR as well as the subjective scores) of the BRISQUE, FRIQUEE, TLVQM, QAWV and NTSSS metrics are the following: frames_per_videox36, frames_per_videox560, 1x75, frames_per_videox32 and NTSSS composed of 3 vectors of 1x4 for the 3DMSCN features and 1x48 for the space-time I and Q features.

The NIQE and VIIDEO metrics being non-mapping metrics, no training (SVR) was performed and the objective scores were correlated with the subjective ones after being treated by the logistic function.

5.4.2 NIQE processing

Spatial-temporal pooling

Many objective VQA algorithms include a key step of spatial-temporal grouping of frame quality scores. The experimental results show that the time grouping method reveals the robustness of the higher performance models, as these make the score given by the NR quality prediction NR metrics more correlated with the subjective studies.

According to [31], when using the time grouping methods (Arithmetic Mean, Harmonic Mean, Geometric Mean, Minkowski Mean, Percentile, VQPooling, Temporal Variation, Primacy Effect, Recency Effect y Temporal Hysteresis) in the NIQE [12] metric, they observe that the best behavior with respect to the KoNViD-1k and LIVE-VQC databases is VQPooling, because this approach provides the PLCC and SROCC closest to 1.

VQPooling: This strategy is an adaptive spatial and temporal pooling strategy proposed in [32]. Here we only study the temporal pooling part, wherein the quality scores of all frames are classified into two groups composed of higher and lower quality, using k-means clustering. The two groups, dubbed the group of low scores (G_L) and a group of high scores (G_H) are then combined to obtain an overall quality prediction on the entire video sequence[31]. Equation 5 shows how to calculate the VQPooling technique.

$$Q = \frac{\sum_{n \in G_L} q_n + w \cdot \sum_{n \in G_H} q_n}{|G_L| + w \cdot |G_H|}$$

Equation 5. Spatial-Temporal grouping strategy (VQPooling).

Where $|G_L|$ and $|G_H|$ denote the cardinality of G_L and G_H, while the weight w is defined as the ratio between the scores in G_L and G_H, as shown in Equation 6.

$$w = (1 - \frac{M_L}{M_H})^2$$

Equation 6. Ratio between the scores in GL and GH.

where M_L and M_H are the average value of the quality scores in set G_L and G_H, respectively.

NSTSS, NIQE and VIIDEO logistics function

In the hope of modeling the non-linearities present in the objective scores extracted from NSTSS, NIQE, VIIDEO and the MOS, we proceed to pass the metric scores through the logistic function of Equation 7 following the procedure described in [47]. Once this was done, the new objective scores were used to obtain the correlation indices presented below.

$$Q'_j = \beta_2 + \frac{\beta_1 - \beta_2}{1 + exp^{-(Qj - \frac{\beta_3}{|\beta_4|})}}$$

Equation 7. Logistic function model.

5.4.3 NSTSS processing

To process this metric we found that it is necessary to have a more robust computer equipment than those available in this work, because when running it we found that the Ram and the available graphic card were not enough, so we processed as much as possible as shown in Table 10, being the percentage extracted from the entire database with the equipment of Table 11, where the latter has similar specifications to those used to perform the time tests in [7], but in which they used a database with QCIF and CIF video resolution.

Extraction			
Database	**SpatioTemporal 3DMSCN**	**SpatioTemporal I**	**SpatioTemporal Q**
LIVE-Qualcomm	0%	0%	0%
LIVE-VQC	27.50%	0%	0%
KoNViD-1k	22.50%	0%	0%
CVD2014	24.30%	7%	16.6%

Table 10. Extraction and Training with NSTSS Metric.

Reference	RAM	GPU
Lenovo G40 I5	8 Gb	AMD RADEON
Acer E5-575G-56GB	16 Gb	Nvidia geforce 940MX
Dell precision 5520	32 Gb	N/A
Servidor CPU I7- 8700K 6 cores	40 Gb	Titan XP 12 GB

Table 11. Available computer resources.

5.4.4 Support vector regressor (SVR)

The SVR is capable of handling high dimensional data, comparable to the feature length vector extracted from the FRIQUEE, TLVQM, QAWV and BRISQUE metrics. Therefore, Matlab's machine learning toolbox was used to implement the SVR with a radial base function (RBF) kernel. 80% of the dataset was utilized for training and the remaining 20% was used for testing. During training process 350 iterations were ran to optimize automatically the SVR hyperparameters that best predict the MOS based on the perceptual quality inputs. We conducted 250 random iterations (i.e shuffle and interchange training and test datasets) and in each iteration, SRCC, PCC and RMSE were calculated against the human scores.

5.4.5 Performance of VQA NR metrics in databases

The performance values of the seven state-of-the-art metrics in four VQA datasets are summarized in Tables 12,13 and 14

VQA Method	LIVE-Qualcomm	LIVE-VQC	KoNViD-1k	CVD2014
BRISQUE	0.5060 ± 0.1190	0.5640 ± 0.0750	0.5930 ± 0.0410	0.4580 ± 0.1120
FRIQUEE	**0.6108 ± 0.1229**	0.6500 ± 0.1545	0.6465 ± 0.1226	0.8105 ± 0.0663
TLVQM	0.5804 ± 0.1876	**0.7793 ± 0.0439**	0.7327 ± 0.0495	0.7240 ± 0.0883
QAWV	0.5400 ± 0.2215	0.7500 ± 0.0755	**0.7823 ± 0.0828**	**0.8303 ± 0.2084**
VQPooling NIQE	0.3799 ± 0.0314	-0.0308 ± 0.0179	-0.2078 ± 0.0175	0.1641 ± 0.0363
Average-pooled NIQE	0.4606 ± 0.0284	-0.0031 ± 0.1026	-0.5276 ± 0.0114	0.312 ± 0.0299
VIIDEO	0.1102 ± 0.0321	0.1038 ± 0.0233	0.3048 ± 0.0138	0.1163 ± 0.0310
3D-MSCN	X	0.2807 ± 0.1591	0.3205 ± 0.3786	0.4502 ± 0.2916

Table 12. Median PLCC ± Standard Deviation of proposed VQA method.The results of the methods were evaluated on the all datasets.

VQA Method	LIVE-Qualcomm	LIVE-VQC	KoNViD-1k	CVD2014
BRISQUE	0.4940 ± 0.1030	0.5790 ± 0.0560	0.6260 ± 0.0380	0.5110 ± 0.1060
FRIQUEE	**0.5789 ± 0.1067**	0.6492 ± 0.0485	0.6575 ± 0.0389	0.7879 ± 0.0628
TLVQM	0.5597 ± 0.1446	**0.7843 ± 0.0376**	0.7469 ± 0.0314	0.7299 ± 0.090
QAWV	0.4760 ± 0.2066	0.6996 ± 0.0845	**0.7757 ± 0.0932**	**0.8614 ± 0.2237**
VQPooling NIQE	0.3756 ± 0.0318	0.2155 ± 0.0204	-0.3064 ± 0.0122	0.3451 ± 0.0329
Average-pooled NIQE	0.4313 ± 0.0289	0.2864 ± 0.020	-0.5355 ± 0.0125	0.4872 ± 0.0321
VIIDEO	0.1381 ± 0.0359	0.0164 ± 0.0201	0.3068 ± 0.0141	0.0911 ± 0.0342
3D-MSCN	X	0.2026 ± 0.1519	0.3096 ± 0.1402	0.4045 ± 0.3243

Table 13. Median SROCC ± Standard Deviation of proposed VQA method.The results of the methods were evaluated on the all datasets.

VQA Method	LIVE-Qualcomm	LIVE-VQC	KoNViD-1k	CVD2014
BRISQUE	10.4450 ± 1.1350	14.0050 ± 0.8470	10.4500 ± 0.5010	19.5380 ± 3.9790
FRIQUEE	**9.7150 ± 1.4730**	12.9500 ± 4.4700	12.5015 ± 1.1811	12.7620 ± 1.8710
TLVQM	10.1660 ± 1.7120	**10.6828 ± 0.9583**	10.9580 ± 0.9238	15.4255 ± 2.1393
QAWV	11.6197 ± 2.2747	12.1337 ± 1.4093	**0.3990 ± 0.0578**	**12.3669 ± 3.7055**
VQPooling NIQE	19.5398 ± 0.3799	22.924 ± 0.3929	0.7404 ± 0.0085	27.614 ± 0.8151
Average-pooled NIQE	21.7613 ± 0.3903	72.1562 ± 23.0766	1.0673 ± 0.0099	51.0431 ± 0.7084
VIIDEO	12.0650 ± 0.2591	17.0255 ± 0.2279	0.6139 ± 0.0063	21.3166 ± 0.4247
3D-MSCN	X	17.9263 ± 2.1769	0.6558 ± 0.1885	18.3050 ± 3.5347

Table 14. Median RMSE ± Standard Deviation of proposed VQA method.The results of the methods were evaluated on the all datasets.

5.4.6 Performance of VQA NR metrics in LIVE-Qualcomm by type of distortion and device

Since the LIVE-Qualcomm data set contains videos classified into 6 authentic distortions, the data set to be trained was organized in such a way that each distortion was represented as follows 20% of the data for each type of distortion for testing and 80% for training, except for QAWV with 20% for validation, 20% for testing and 60% for training. The results of the performance of the VQA NR metrics are shown in Tables 15, 16 and 17.

VQA Method	Artifacts	Color	Exposure	Focus	Sharpness	Stabilization	All distortions
BRISQUE	0.4177 ± 0.2951	0.4131 ± 0.328	0.4769 ± 0.2931	0.7295 ± 0.2546	0.574 ± 0.2825	0.4021 ± 0.4027	0.5518 ± 0.0974
FRIQUEE	0.7086 ± 0.2300	0.312 ± 0.3850	**0.6913 ± 0.2626**	0.6471 ± 0.3683	0.6556 ± 0.2926	0.6783 ± 0.3313	**0.6179 ± 0.1056**
TLVQM	0.6122 ± 0.2513	**0.4992 ± 0.2201**	0.4396 ± 0.2063	**0.8027 ± 0.2003**	**0.8263 ± 0.1981**	**0.7619 ± 0.2208**	0.6177 ± 0.1409
QAWV	**0.7530 ± 0.2048**	0.3552 ± 0.1828	0.5364 ± 0.3645	0.4024 ± 0.2757	0.7671 ± 0.0619	0.2975 ± 0.1116	0.5600 ± 0.2715
VQPooling NIQE	0.2426 ± 0.0582	-0.0344 ± 0.1282	0.0893 ± 0.0968	0.4693 ± 0.0809	0.5281 ± 0.0762	0.355 ± 0.0718	0.3799 ± 0.0314
Average-pooled NIQE	0.5391 ± 0.0721	0.1488 ± 0.1928	0.106 ± 0.0492	0.4901 ± 0.073	0.3885 ± 0.1232	0.5003 ± 0.0958	0.4606 ± 0.0284
VIIDEO	0.012 ± 0.0903	-0.0861 ± 0.142	0.1857 ± 0.0994	0.1791 ± 0.0834	0.1645 ± 0.0776	-0.1567 ± 0.075	0.1102 ± 0.0321

Table 15. Median PLCC ± Standard Deviation of proposed VQA method.The results of the methods were evaluated on the same dataset (LIVE-Qualcomm).

VQA Method	Artifacts	Color	Exposure	Focus	Sharpness	Stabilization	All distortions
BRISQUE	0.4290 ± 0.3300	0.3210 ± 0.3140	0.3930 ± 0.3210	0.7500 ± 0.2720	0.5360 ± 0.2950	0.3570 ± 0.3680	0.5090 ± 0.1030
FRIQUEE	0.6786 ± 0.2419	0.335 ± 0.3978	**0.6071 ± 0.2985**	0.6571 ± 0.3930	0.6429 ± 0.2784	0.6071 ± 0.2997	**0.5805 ± 0.1025**
TLVQM	0.5714 ± 0.2241	0.4286 ± 0.2561	0.3929 ± 0.2143	**0.7500 ± 0.2295**	**0.7857 ± 0.2233**	**0.6786 ± 0.1808**	0.5737 ± 0.1484
QAWV	**0.7727 ± 0.2236**	**0.4286 ± 0.1117**	0.5238 ± 0.3295	0.3939 ± 0.2029	0.6429 ± 0.0905	0.4048 ± 0.1027	0.506 ± 0.2470
VQPooling NIQE	0.2876 ± 0.0776	-0.179 ± 0.1117	0.1117 ± 0.0968	0.3562 ± 0.1052	0.3892 ± 0.0932	0.3041 ± 0.083	0.3756 ± 0.0318
Average-pooled NIQE	0.3826 ± 0.0867	-0.1053 ± 0.1192	0.2884 ± 0.0823	0.4246 ± 0.0818	0.5775 ± 0.0717	0.4095 ± 0.0737	0.4313 ± 0.0289
VIIDEO	-0.099 ± 0.0779	-0.1107 ± 0.1299	0.38 ± 0.083	0.0765 ± 0.0928	0.1932 ± 0.1014	-0.0632 ± 0.0908	0.1381 ± 0.0359

Table 16. Median SROCC ± Standard Deviation of proposed VQA method.The results of the methods were evaluated on the same dataset (LIVE-Qualcomm).

VQA Method	Artifacts	Color	Exposure	Focus	Sharpness	Stabilization	All distortions
BRISQUE	11.5410 ± 2.8360	**8.3600 ± 1.9440**	10.3390 ± 2.349	**9.6610 ± 2.8450**	9.7060 ± 2.1110	8.5540 ± 2.6090	10.0080 ± 0.9250
FRIQUEE	9.3230 ± 2.5050	10.5390 ± 2.3210	9.2610 ± 2.2500	11.0690 ± 3.1320	8.5310 ± 2.2810	7.4520 ± 2.4520	**9.8350 ± 97289**
TLVQM	10.9150 ± 3.7840	9.4780 ± 2.9300	11.8430 ± 2.5440	9.8290 ± 3.3420	**8.3530 ± 2.7280**	7.3130 ± 2.0190	10.2060 ± 1.5670
QAWV	**8.5378 ± 1.3725**	10.3738 ± 1.2946	**7.9002 ± 1.6594**	13.574 ± 4.1126	9.3686 ± 1.0933	**7.2965 ± 1.0543**	14.4197 ± 2.3647
VQPooling NIQE	18.5909 ± 1.3763	14.9977 ± 1.0532	21.0495 ± 1.4606	15.9568 ± 1.2452	13.0289 ± 0.8736	18.1224 ± 1.4933	19.5398 ± 0.3799
Average-pooling NIQE	20.8613 ± 1.2765	18.9328 ± 0.7685	25.8699 ± 1.1595	21.3108 ± 1.1905	19.9309 ± 1.1202	20.2597 ± 0.6607	21.7613 ± 0.3903
VIIDEO	12.8631 ± 0.5798	13.2405 ± 0.0861	12.5237 ± 0.5433	13.3822 ± 0.7004	14.0196 ± 0.4885	12.5821 ± 0.6533	12.065 ± 0.2591

Table 17. Median RMSE ± Standard Deviation of proposed VQA method.The results of the methods were evaluated on the same dataset (LIVE-Qualcomm).

On the other hand, LIVE-Qualcomm contains videos acquired from 8 different smartphones for each unique scene. To avoid redundancy of information in the performance of the SVR, duplicate videos of the same scene were eliminated. Therefore, the data set was organized so that each device was represented by the same number of videos (Galaxy GS5=8, Galaxy GS6=8, HTC One VX=8, Iphone 5S=8, LG G2=8, Lumia 1020 = 3, Note 4 =4, Oppo find 7 =7) obtaining a total of 54 videos [49]. Tables 18, 19 and 20 tabulate the results of the performance of the VQA NR metrics.

VQA Method	Samsung Galaxy S5	Samsung Galaxy S6	HTC One VX	Apple Iphone 5S	LG G2	Nokia Lumia 1020	Samsung Galaxy Note 4	Oppo Find 7	All devices
BRISQUE	0.5355 ± 0.4530	0.3616 ± 0.3578	0.3668 ± 0.3840	0.7460 ± 0.5241	0.4910 ± 0.3392	0.3491 ± 0.6896	0.5183 ± 0.5325	0.0865 ± 0.3623	0.5893 ± 0.2295
FRIQUEE	0.6203 ± 0.6430	**0.7316 ± 0.3778**	0.5268 ± 0.4540	0.7560 ± 0.6441	**0.5710 ± 0.3292**	0.5191 ± 0.6296	0.6383 ± 0.5125	0.4350 ± 0.5123	0.5377 ± 0.2074
TLVQM	**0.7568 ± 0.5957**	0.5901 ± 0.4830	0.3807 ± 0.3980	**0.8888 ± 0.2550**	0.3103 ± 0.4234	0.4651 ± 0.6798	**0.7672 ± 0.4632**	0.2321 ± 0.4804	**0.7431 ± 0.2086**
QAWV	0.1628 ± 0.4470	0.6319 ± 0.4597	0.5687 ± 0.2446	0.8692 ± 0.1211	0.5184 ± 0.2816	**0.6635 ± 0.0847**	0.2311 ± 0.1139	**0.7101 ± 0.5393**	0.5800 ± 0.2349
VQPooling NIQE	0.2588 ± 0.0772	0.3343 ± 0.0683	0.4262 ± 0.0917	0.616 ± 0.0744	0.2455 ± 0.1281	0.2887 ± 0.1112	0.4117 ± 0.073	0.2144 ± 0.0728	0.4156 ± 0.0342
Average-pooled NIQE	0.2898 ± 0.0834	0.3323 ± 0.0751	0.5541 ± 0.063	0.544 ± 0.0843	0.3616 ± 0.0917	-0.2552 ± 0.3154	0.5976 ± 0.0627	0.5409 ± 0.0774	0.1214 ± 0.0387
VIIDEO	0.0147 ± 0.2314	-0.0035 ± 0.0489	**0.6994 ± 0.0453**	-0.0297 ± 0.1121	0.163 ± 0.0817	-0.1418 ± 0.0693	0.1986 ± 0.1247	0.3177 ± 0.1299	0.1812 ± 0.033

Table 18. Median PLCC ± Standard Deviation of proposed VQA method.The results of the methods were evaluated on the same dataset (Qualcomm).

VQA Method	Samsung Galaxy S5	Samsung Galaxy S6	HTC One VX	Apple Iphone 5S	LG G2	Nokia Lumia 1020	Samsung Galaxy Note 4	Oppo Find 7	All devices
BRISQUE	0.4000 ± 0.4580	0.3571 ± 0.3602	0.3714 ± 0.3966	0.8000 ± 0.5071	0.4857 ± 0.3763	0.5000 ± 0.7244	0.4000 ± 0.5194	0.0286 ± 0.3984	0.5864 ± 0.2232
FRIQUEE	**0.5970 ± 0.6180**	**0.7271 ± 0.3602**	0.5043 ± 0.4266	0.7352 ± 0.6071	**0.5357 ± 0.3763**	0.4984 ± 0.5844	0.6241 ± 0.5194	0.3840 ± 0.4951	0.5137 ± 0.3590
TLVQM	0.4000 ± 0.5255	0.4857 ± 0.4242	0.4857 ± 0.3837	**0.8000 ± 0.2781**	0.2857 ± 0.4254	0.5000 ± 0.6670	**0.8000 ± 0.4663**	0.2857 ± 0.4534	**0.7046 ± 0.2052**
QAWV	0.5000 ± 0.5705	0.6666 ± 0.4181	0.5000 ± 0.2672	0.6500 ± 0.1172	0.3818 ± 0.2672	**0.7000 ± 0.0805**	0.1786 ± 0.1417	**0.5250 ± 0.5273**	0.4960 ± 0.2323
VQPooling NIQE	0.3581 ± 0.1028	0.2908 ± 0.0891	0.3755 ± 0.0959	0.6846 ± 0.0873	0.1554 ± 0.0949	0.1392 ± 0.1337	0.5353 ± 0.0807	0.1758 ± 0.1179	0.3798 ± 0.0286
Average-pooled NIQE	0.3758 ± 0.0962	0.4111 ± 0.0932	0.4518 ± 0.079	0.5678 ± 0.1118	0.2037 ± 0.0912	0.2538 ± 0.2028	0.4853 ± 0.0939	0.2433 ± 0.0931	0.1726 ± 0.038
VIIDEO	0.1863 ± 0.1366	0.2274 ± 0.1037	**0.6194 ± 0.0636**	-0.1604 ± 0.1375	0.345 ± 0.0717	-0.0734 ± 0.1735	0.0062 ± 0.1224	0.3948 ± 0.0843	0.1656 ± 0.0331

Table 19. Median SROCC ± Standard Deviation of proposed VQA method.The results of the methods were evaluated on the same dataset (Qualcomm).

VQA Method	Samsung Galaxy S5	Samsung Galaxy S6	HTC One VX	Apple Iphone 5S	LG G2	Nokia Lumia 1020	Samsung Galaxy Note 4	Oppo Find 7	All devices
BRISQUE	8.0993 ± 2.6390	10.1976 ± 2.2885	14.2224 ± 2.7736	10.3809 ± 2.8149	8.4309 ± 2.2771	9.8677 ± 5.5404	10.0416 ± 3.0672	10.0661 ± 2.7636	9.8002 ± 1.9968
FRIQUEE	8.8530 ± 3.7369	**8.4618 ± 5.8141**	10.2060 ± 4.7414	10.0000 ± 5.2019	**8.1342 ± 3.7468**	**9.6075 ± 12.2930**	9.0602 ± 6.9392	10.128 ± 7.9075	11.3010 ± 2.2449
TLVQM	7.3005 ± 7.5952	9.2072 ± 2.8719	13.0060 ± 1.7017	**7.5709 ± 2.8485**	9.4152 ± 3.1761	10.9140 ± 4.5148	**8.4116 ± 2.0898**	9.4318 ± 3.6503	**8.6683 ± 3.0326**
QAWV	**4.6404 ± 2.3863**	11.7063 ± 4.1553	**7.8857 ± 1.5310**	7.7245 ± 1.6291	13.1268 ± 4.5048	10.3327 ± 1.3008	9.0973 ± 0.8471	10.7772 ± 2.7262	13.5597 ± 2.3447
VQPooling NIQE	10.941 ± 0.5634	11.3904 ± 1.0944	8.8865 ± 0.5735	12.6669 ± 1.0803	10.7673 ± 0.936	25.8833 ± 1.384	10.5937 ± 0.6508	11.0426 ± 1.1773	11.4259 ± 0.2802
Average-pooled NIQE	9.2822 ± 0.7811	11.1229 ± 1.1086	9.5627 ± 0.4938	12.9311 ± 1.2404	10.3468 ± 0.6989	12.8961 ± 1.4029	10.6292 ± 0.7766	**8.3124 ± 0.5372**	11.8962 ± 0.343
VIIDEO	76.0245 ± 17.1676	12.2814 ± 1.2175	9.1265 ± 0.7382	20.9853 ± 1.1363	11.7482 ± 0.8945	16.1526 ± 2.0654	14.7634 ± 1.1758	9.6715 ± 0.7306	12.2962 ± 0.3603

Table 20. Median RMSE ± Standard Deviation of proposed VQA method.The results of the methods were evaluated on the same dataset (Qualcomm).

5.4.7 Performance of VQA NR metrics in CVD2014 by scene type

Since the CVD2014 data set contains videos classified in 5 scenes, the data set to be trained was organized in such a way that each scene was represented as follows: 20% of the data of each type of scene for testing and 80% for training, except for QAWV with 20% for validation, 20% for testing and 60% for training. Performance results for VQA NR metrics are shown in Tables 21, 22 and 23.

VQA Method	City	Newspaper	Talkinghead	Television	Traffic	ALL
BRISQUE	0.6324 ± 0.1479	0.7234 ± 0.4465	0.3239 ± 0.178	0.283 ± 0.3488	0.5518 ± 0.5013	0.4796 ± 0.1201
FRIQUEE	0.8037 ± 0.1435	0.8821 ± 0.2763	0.7722 ± 0.136	0.8767 ± 0.1387	0.7817 ± 0.3004	0.8005 ± 0.0686
TLVQM	0.8361 ± 0.0911	0.7251 ± 0.3666	0.6224 ± 0.201	0.8297 ± 0.1875	**0.8639 ± 0.3213**	0.7443 ± 0.0699
QAWV	**0.9075 ± 0.1526**	**0.9224 ± 0.018**	**0.864 ± 0.1839**	**0.9205 ± 0.041**	0.6589 ± 0.2493	**0.8538 ± 0.2188**
VQPooling NIQE	0.3856 ± 0.0632	0.1012 ± 0.1217	0.2661 ± 0.0499	0.0711 ± 0.1035	0.2705 ± 0.123	0.1641 ± 0.0363
Average-pooled NIQE	0.5887 ± 0.0416	0.6322 ± 0.054	0.4243 ± 0.0498	0.5113 ± 0.0535	0.1289 ± 0.121	0.312 ± 0.0299
VIIDEO	-0.1256 ± 0.097	0.1635 ± 0.1053	0.0703 ± 0.0529	-0.3824 ± 0.1003	0.4035 ± 0.932	0.1163 ± 0.031
3D-MSCN	0.3019 ± 0.5805	X	0.2203 ± 0.5445	X	0.4340 ± 0.5251	0.3532 ± 0.2641

Table 21. Median PLCC ± Standard Deviation of proposed VQA method.The results of the methods were evaluated on the same dataset (CVD2014).

VQA Method	City	Newspaper	Talkinghead	Television	Traffic	ALL
BRISQUE	0.6103 ± 0.1677	0.6000 ± 0.4465	0.3882 ± 0.1699	0.2571 ± 0.3846	0.4000 ± 0.5426	0.5035 ± 0.1045
FRIQUEE	0.7393 ± 0.1375	0.8000 ± 0.2803	**0.7198 ± 0.1557**	0.8286 ± 0.1900	**0.8000 ± 0.3295**	0.7935 ± 0.0676
TLVQM	0.7910 ± 0.094	0.6000 ± 0.4150	0.5440 ± 0.1990	0.7710 ± 0.2340	0.7000 ± 0.3800	0.7370 ± 0.0760
QAWV	**0.8570 ± 0.1570**	**0.9330 ± 0.0690**	0.6000 ± 0.1410	**0.9310 ± 0.0860**	0.5800 ± 0.1930	**0.8650 ± 0.2360**
VQPooling NIQE	0.4524 ± 0.0510	0.1671 ± 0.1032	0.2356 ± 0.0590	0.0277 ± 0.1083	0.1607 ± 0.1319	0.3451 ± 0.0329
Average-pooled NIQE	0.5493 ± 0.0433	0.6155 ± 0.0671	0.3900 ± 0.0498	0.3446 ± 0.0751	-0.0464 ± 0.1392	0.4872 ± 0.0321
VIIDEO	-0.0008 ± 0.0751	0.0661 ± 0.1031	0.0892 ± 0.0565	-0.4458 ± 0.084	0.3625 ± 0.1093	0.0911 ± 0.0342
3D-MSCN	0.3000 ± 0.5661	X	0.2000 ± 0.5639	X	0.3210 ± 0.5948	0.2909 ± 0.2794

Table 22. Median SROCC ± Standard Deviation of proposed VQA method.The results of the methods were evaluated on the same dataset (CVD2014).

VQA Method	City	Newspaper	Talkinghead	Television	Traffic	ALL
BRISQUE	18.8820 ± 3.2500	17.3350 ± 6.0768	21.7580 ± 15.0030	15.6540 ± 4.3597	16.5920 ± 5.6713	19.7570 ± 7.5555
FRIQUEE	**12.4060 ± 4.3629**	**11.0350 ± 3.8290**	13.7150 ± 3.0674	8.0408 ± 3.1729	12.8330 ± 3.4094	12.8190 ± 2.1651
TLVQM	13.1900 ± 2.5000	15.7810 ± 3.8117	16.7880 ± 3.2000	7.7278 ± 2.8811	**11.9590 ± 4.3095**	14.6560 ± 1.8706
QAWV	16.8779 ± 6.2343	11.9417 ± 2.4273	**8.9842 ± 2.8612**	**4.9127 ± 3.5369**	13.5354 ± 2.5705	**12.4679 ± 3.8455**
VQPooling NIQE	26.3566 ± 0.8358	23.5695 ± 1.5437	23.2622 ± 0.8836	18.2013 ± 1.5331	17.8906 ± 1.6524	27.614 ± 0.8151
Average-pooled NIQE	27.5661 ± 1.1097	27.2745 ± 1.6094	31.0655 ± 0.9792	33.6677 ± 1.2513	35.1797 ± 2.0916	51.0431 ± 0.7084
VIIDEO	25.4152 ± 1.2738	25.7376 ± 1.705	26.066 ± 1.6302	31.6792 ± 1.4398	17.2354 ± 1.0226	21.3166 ± 0.4247
3D-MSCN	21.0100 ± 9.0493	X	16.5120 ± 5.3688	X	18.6900 ± 5.2395	19.3540 ± 4.4823

Table 23. Median RMSE ± Standard Deviation of proposed VQA method.The results of the methods were evaluated on the same dataset (CVD2014).

CHAPTER 6

Analysis of results

Traditional performance analysis is measured by comparing the predictions of the algorithm with the values of the MOS.

Firstly analyzing the results presented in Tables 12,13 and 14, the TLVQM model exceeds all the benchmarks in the LIVE-VQC dataset, this is because this predictor focuses mainly on the motion characteristics of the videos, being this database rich in this aspect, so the perceptive characteristics of content and the modelling of the effects of the temporal memory are important for the evaluation of the quality of no-reference video; however, the difference with the second best QAWV method is not very significant, being the latter superior in KoNViD-1k and CVD2014. On the other hand, VIIDEO and NIQE results correlate poorly with MOS, while for the other models the regular trend against MOS is followed.

Secondly, when analyzing Tables 18, 19 and 20 corresponding to the study of performance by device, it is observed that videos recorded by Iphone 5s, followed by the Samsung GS6 obtained the highest correlation rates with TLVQM, FRIQUEE and QAWV metrics, since these devices have a built-in optical image stabilizer, as well as better recording quality (resolution) than the other devices studied [48], which suggests that VQA NR metrics have made a significant advance in prediction.

Thirdly, we found some remarkable categories of distortion in which the blind models considered had a particularly bad or good performance (Tables 14, 15 and 16). For example, FRIQUEE and QAWV correlated very well with videos categorized with artifact and exposure distortions, TLVQM performed best in focus, dark and stabilization distortions, but all metrics correlated poorly with videos categorized with color distortions, which is probably due to the fact that most metrics extract luminance characteristics based on NSS, driven by perception.

It is also observed that the average temporal clustering of NIQE scores correlates better with subjective opinions compared to VQPooling. On the other hand, the worst performance was obtained by the VIIDEO metric since this model depends only on the temporal scene statistics to make quality predictions and the videos in the studied databases contain

rich content in authentic mixtures of spatial distortions, besides movement and ego, which makes this mixture of distortions difficult for VIIDEO.

Following with the previous analysis, Tables 21, 22 and 23 show the performance of the metrics in terms of PLCC, SROCC and RMSE for the CVD2014 database by scene. According to the results, QAWV had the highest performance in terms of prediction. The second place was obtained by FRIQUEE and the third place by TLVQM. For which it is observed that the correlation rates are the highest obtained in this study and this is because sharpness is the most important quality dimension when describing the quality of CVD2014 videos [36]. Furthermore, it should be noted that the videos in this database contain audio and in the subjective experiments the participants were asked to rate the overall quality and not the visual quality and we all know that most of the real life videos include audio and their effect on the perception of the overall quality is evident [36].

It is observed that when training by device, distortion and scene the correlation rates increased since there is an equity in the data when training by emphasizing the rating of the videos improving the prediction for videos with mixed authentic distortions.

In addition to performance, computational efficiency is also crucial for NR-VQA methods.

- In Tables 6 and 7 it was found that for the tests carried out for BRISQUE and NIQE a reduction in computation time of 24s and 0.8s respectively was obtained when processing in Python versus Matlab, but a higher standard deviation of 27.7 and 13.5 tends to be obtained, that is because it was determined to use the official version of the metrics implemented in Matlab.
- In Table 8, it can be seen that FRIQUEE is the metric with the longest execution time, this is because it extracts the characteristics of the video through 4 layers and then performs the prediction based on a SVR. The dimension of the feature map for each frame is 560. In contrast, QAWV, TLVQM, BRISQUIE and NIQE metrics reduce the dimensions of the feature maps achieving a faster calculation between vectors, In addition, they rely on software such as CUDA, MPI and Parallel Computing Toolbox.

CHAPTER 7

Conclusion

In this document, we have evaluated 7 state-of-the-art VQA NR metrics, training and validating the proposed methods using four relevant video quality databases with a wide variety of video content specifically targeted at user-generated content that is prone to capture artifacts, such as camera shake, over and under-exposure, among other authentic distortions.

The results show that in terms of accuracy, TLVQM, QAWV and FRIQUEE methods outperform all reference methods tested. The computational advantage of TLVQM is significant given its execution time and efficiency in the calculation of objective scores.

7.1 Future work

Future work could involve a faster implementation of the VQ NR algorithms, as well as improving their execution times so that they are useful for real-time video applications, considering the optimization done in FRIQUEE could help solve the problem of resource allocation for future research.

We will consider that VQA NR models focused on spatial-temporal models as well as on motion characteristics should be further developed since the best performances regarding correlation with MOS were of metrics that model these features. However, the metrics evaluated in this research performed poorly on color distortion, so there is ample room to develop an objective model that correlates well with human perception from this approach.

Finally, we suggest a more rigorous study of average pooling and VQPooling for NIQE because the expected results were not obtained. Also use more computer equipment to finish the study with the NSTSS metric, in which it is expected that by using the 3DMSCN and SpatioTemporal I and Q features together, a much better result will be obtained than each one independently.

CHAPTER 8

References

[1] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda and K. Yang, "In-Capture Mobile Video Distortions: A Study of Subjective Behavior and Objective Algorithms," *:in IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2061-2077, Sept. 2018.

[2] Z. Sinno and A. C. Bovik, "Large-Scale Study of Perceptual Video Quality," *in IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612-627, Feb. 2019.

[3] D. Ghadiyaram and A.C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," J. of Vision, https://arxiv.org/abs/1609.04757.

[4] Li, D., Jiang, T., Jiang, M. (2019, October). Quality assessment of in-the-wild videos. In Proceedings of the 27th ACM International Conference on Multimedia (pp. 2351-2359).

[5] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," *in IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695-4708, Dec. 2012.

[6] Mittal, A., Soundararajan, R., Bovik, A. C. (2012). Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3), 209-212.

[7] Dendi, S. V. R., Channappayya, S. S. (2020). No-Reference Video Quality Assessment Using Natural Spatiotemporal Scene Statistics. *:*IEEE Transactions on Image Processing, 29, 5612-5624.

[8] Korhonen, J. (2019). Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12), 5923-5938.

[9] M. Shahid, A. Rossholm, B. Lövström, and H.-J. Zepernick, "No-reference image and video quality assessment: a classification and review of recent approaches," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, 2014.

[10] "Cisco Study Reveals 80% of the World's Internet Traffic Will Be Video By 2019 - Purposeful Films", Purposeful Films, 2020. [Online]. Available: https://www.purposefulfilms.com/cisco-study-reveals-80-of-the-worlds-internet-traffic-will-be-video-by-2019/. [Accessed: 01- Feb- 2020].

[11] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," Visual Communications and Image Processing 2003, 2003.

[12] D. Ghadiyaram, J. Pan, A. C. Bovik, A. Moorthy, P. Panda and K. Yang, "Subjective and objective quality assessment of Mobile Videos with In-Capture distortions," 2017 *IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 1393-1397.
[13] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives," Multimedia Systems, vol. 22, no. 2, pp. 213–227, 2014.
[14] Q. Huynh-Thu, M. Garcia, F. Speranza, P. Corriveau and A. Raake, "a regressor for each no-reference video quality model to predict human scores of perceptual video quality," *in IEEE Transactions on Broadcasting*, vol. 57, no. 1, pp. 1-14, March 2011.
[15] Z. Wang, "Applications of Objective Image Quality Assessment Methods [Applications Corner]," *in IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 137-142, Nov. 2011.
[16] D. Ghadiyaram and A.C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, 2016.
[17] D. Ghadiyaram and A.C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," J. of Vision, https://arxiv.org/abs/1609.04757.
[18] B. Ortiz-Jaramillo, J. Niño-Castañeda, L. Platiša, and W. Philips, "Content-aware objective video quality assessment," *Journal of Electronic Imaging*, vol. 25, no. 1, p. 013011, 2016.
[19] D. Li, T. Jiang, W. Lin, and M. Jiang, "Which Has Better Visual Quality: The Clear Blue Sky or a Blurry Animal?," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1221–1234, 2019.
[20] Ernestasia Siahaan,Alan Hanjalic,and Judith A Redi.2018. Semantic-aware blind image quality assessment. SPIC 60(2018),237–252.
[21] Jin jian Wu,Jichen Zeng,Weisheng Dong,Guangming Shi,and Weisi Lin.2019. Blind image quality assessment with hierarchy:Degradation from local structure to deep semantics. JVCIR 58(2019),353–362.
[22] "SPSS Tutorials: Pearson Correlation," LibGuides. [Online]. Available: https://libguides.library.kent.edu/SPSS/PearsonCorr. [Accessed: 30-Apr-2020].
[23] Basic Concepts of Correlation — Real Statistics Using Excel." [Online]. Available: http://www.real-statistics.com/correlation/basic-concepts-correlation/. [Accessed: 30-Apr-2020].
[24] A. Lehman, JMP for basic univariate and multivariate statistics. Cary, NC: SAS Press, 2005, p. 123.
[25] Myers and A. Well, Research design and statistical analysis, 2nd ed. Mahwah, N.J.: Lawrence Erlbaum Associates, 2003, p. 508
[26] "The Concise Encyclopedia of Statistics," 2008.
[27] A. Al Jaber, and H. Elayyan, "Toward quality assurance and excellence in higher education," *River Publishers*, p. 284, 2018.

[28]E. Lehmann, and G. Casella, *Theory of point estimation*, 2nd. ed. New York: Springer, 1998.
[29]"Laboratory for Image and Video Engineering - The University of Texas at Austin", Live.ece.utexas.edu, 2020. [Online]. Available: https://live.ece.utexas.edu/. [Accessed: 30- Apr- 2020].
[30] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Sziranyi, S. Li, and D. Saupe, "The Konstanz natural video database (KoNViD-1k)," 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), 2017.
[31] N. D. Narvekar, and L. J. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2678–2683, 2011.
[32] D. Hasler, and S. E. Suesstrunk, "Measuring colorfulness in natural images," in Human vision and electronic imaging VIII, vol. 5007, pp. 87–95, 2003.
[33] E. Peli, "Contrast in complex images," Journal of the Optical Society of America, A, vol. 7, no. 10, pp. 2032–2040, 1990.
[34] ITU-T, "Subjective video quality assessment methods for multimedia applications," ITU-T Recommendation P.910, 2008.
[35] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.
[36] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "CVD2014—A database for evaluating no-reference video quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3073-3086, 2016.
[37] A. C. Bovik, "Automatic Prediction of Perceptual Image and Video Quality," in Proceedings 26 of the IEEE, vol. 101, no. 9, pp. 2008-2024, September 2013.
[38] K. Seshadrinathan, and A. C. Bovik, "Temporal hysteresis model of time varying subjective video quality," in ICASSP, IEEE, pp. 1153–1156, 2011.
[39] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in CVPR, IEEE, pp. 248–255, 2009.
[40] S. T. Hammett, "Motion blur and motion sharpening in the human visual system," in Vision Research, vol. 37, no. 18, pp. 2505-10, September 1997.
[41] "Capacitar y utilizar un modelo de evaluación de la calidad sin referencia- MATLAB Simulink- MathWorks América Latina", La.mathworks.com, 2020. [Online]. Available: https://la.mathworks.com/help/images/
train-and-use-a-no-reference-quality-assessment-model.html. [Accessed: 25- Nov- 2020].
[42] D. L. Ruderman, "The statistics of natural images," Netw., Comput. Neural Syst. vol. 5, no. 4, pp. 517–548, 1994.
[43] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle,"

IEEE Transactions on Image Processing, vol. 25, no. 1, pp. 289-300, 2015.
[44] Ijirset.com, 2020. [Online]. Available: https://www.ijirset.com/upload/2018/may/206_Quality_IEEE.pdf . [Accessed: 21- Dec- 2020].
[45] Z. Tu, C. J. Chen, L. H. Chen, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "A Comparative Evaluation of Temporal Pooling Methods for Blind Video Quality Assessment," arXiv preprint arXiv:2002.10651, 2020.
[46] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, "Video Quality Pooling Adaptive to Perceptual Distortion Severity," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 610–620, 2013.
[47] D. E. Moreno-Villamarín, H. D. Benítez-Restrepo, and A. C. Bovik, "Predicting the quality of fused long wave infrared and visible light images," *IEEE Transactions on Image Processing*, vol 26, no. 7, pp. 3479-3491, 2017.
[48] "Apple iPhone 5S vs Samsung Galaxy S6: ¿cuál es la diferencia?", VERSUS, 2020. [Online]. Available: https://versus.com/es/apple-iphone-5s-vs-samsung-galaxy-s6group_cameras. [Accessed: 23- Dec- 2020].
[49] R. G. Nieto, H. D. Benítez-Restrepo, R. F. Quintero, and A. C. Bovik, "No Reference Video Quality Assessment with authentic distortions using 3-D Deep Convolutional Neural Network," Electronic Imaging, vol. 2020, no. 9, 2020.