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

Undergraduate research proposal

José Alejandro Ledesma Mazuera
Stidl Alfonso Torres Morón

Supervised by

Hernán Darío Benítez Restrepo
Roger Alfonso Gómez Nieto

Santiago de Cali, Jun 12, 2020

Resumen

Los sistemas de hardware y procesamiento de videos digitales pueden introducir distorsiones en la señal de video durante el proceso de captura. Estas distorsiones son un factor clave en el éxito de un sistema o servicio multimedia, el cual tiene como propósito lograr que la calidad de la experiencia percibida por el usuario resulte aceptable. Es por esto que en los últimos años se ha acelerado considerablemente el estudio y desarrollo de métodos objetivos automáticos que cuantifican con precisión el impacto de las distorsiones visuales en la percepción sin tener como referencia el video original. La verificación de algoritmos no-referenciados de calidad de vídeo requiere bases de datos realistas de vídeos distorsionados y juicios humanos de los mismos. Sin embargo, la mayoría de las actuales bases de datos de calidad de vídeo disponibles al público se han creado en condiciones altamente controladas utilizando distorsiones simuladas (artificiales) y posteriores a la captura en vídeos de alta calidad. Esta situación motiva a realizar esta propuesta de proyecto de grado, en el cual se propone evaluar las métricas no-referenciadas de última generación tales como FRIQUEE, QAWV, BRISQUE, NIQE, NSTSS y TLVQM en bases de datos de vídeos auténticamente distorsionados como KoNVid-1K, LIVE Qualcomm y LIVE Video Quality Challenge (VQC). Además de proponer el entrenamiento de un regresor, todo esto con el fin de estudiar cómo las distorsiones en la captura son predichas por modelos automáticos de calidad perceptual y evaluadas de acuerdo con la correlación entre estas predicciones y las evaluaciones humanas de calidad.

Palabras clave— Evaluación de calidad de video (VQA), distorsiones en la captura, evaluación subjetiva de la calidad, algoritmos objetivos no-referenciados, distorsiones auténticas, bases de datos, características del contenido, regresión, agrupación espacial - temporal, Coeficiente de Correlación Lineal de Pearson (PLCC), Coeficiente de Correlación de Orden de Rango de Spearman (SROCC) y Error de la Media Cuadrática (RMSE).

Abstract

Digital video processing and hardware systems can introduce certain distortions into the video signal during the capture process. These distortions may affect the success of a multimedia system or service, which aims to make the quality of the experience perceived by the user acceptable. That is why in recent years the study and development of automatic objective methods that accurately quantify the impact of visual distortions in the perception without having the original video as reference has accelerated considerably. Verification of no-reference video quality algorithms requires realistic databases of distorted video and human judgments of it. 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 to carry out this undergraduate project proposal, in which it is proposed to evaluate the latest generation of 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 to proposing the training of a regressor, all this in order to study how distortions in video capture are predicted by automatic perceptual quality models and evaluated according to the correlation between these predictions and human quality assessments.

***Index terms*—-** Video quality assessment (VQA), capture distortions, subjective quality assessment, objective no-reference algorithms, authentic distortions, databases, content-aware features, regression, spatial-temporal pooling, Pearson's Linear Correlation Coefficient (PLCC), Spearman's Rank-Order Correlation Coefficient (SROCC) and Root Mean Square Error (RMSE).

Contents

1 Introduction **4**

2 Research problem **5**
2.1 Problem Statement . . . 5
2.2 Scope of the work . . . 6

3 Objectives **7**
3.1 General Objective . . . 7
3.2 Specific Objectives . . . 7

4 Justification **8**

5 Theoretical framework **9**
5.1 State of the art . . . 9
5.1.1 Subjective methods . . . 9
5.1.2 Objective methods . . . 9
5.1.3 Authentic, In-capture Video Distortions . . . 9
5.1.4 Databases . . . 11
5.1.5 Algorithms . . . 11
5.1.6 Content-Aware Features . . . 13
5.1.7 Regression . . . 13
5.1.8 Spatial-temporal pooling . . . 14
5.1.9 Data analysis measures . . . 14
5.1.10 Hypotheses considered . . . 16
5.1.11 Related work . . . 16

6 Research Methodology **19**
6.1 Description . . . 19
6.1.1 Stage 1: Extracting scores and features from NR VQA metrics . . . 19
6.1.2 Stage 2: Train regressor . . . 19
6.1.3 Stage 3: Performance analysis of NR VQA metrics . . . 20
6.1.4 Stage 4: Writing and reporting results. . . . 20
6.2 Tests . . . 21
6.3 Expected Result . . . 22

7 Resources **23**
7.1 Human . . . 23
7.1.1 Director . . . 23
7.1.2 Co-director . . . 23
7.1.3 Research group of the faculty that supports it. . . . 23
7.2 Economical . . . 24
7.2.1 General resource's budget required . . . 24

8 Schedule **25**

9 References **26**

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].

This document outlines the problem, objectives and rationale behind the need to carry out a no-reference video quality prediction study, and then presents the theoretical framework, methodology, human and physical resources, and activities scheduled to develop this project.

2 Research problem

2.1 Problem Statement

Video traffic already represents 80% of all the mobile internet traffic [3], these multimedia contents are distributed through the telecommunications networks experiencing various types of distortions or degradations 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 [4]. 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 [5]. 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, so they are far from reality and their contributions to the following applications are limited:

1. Monitoring: service providers do not know the quality of the video applications used by their users and do not have objective video quality models that allow them to systematize measurements and monitor the status of the network in a simple and controlled manner [6].
2. Quality control and management: when users pay for video applications of a certain quality, it is, therefore, necessary to measure the quality perceived by all users and estimate how it will be affected by the inclusion of new users in the system [7].
3. Quality based pricing: some services do not have differential pricing based on video quality, due to the inability to measure perceived quality and control the quality that the user is paying [8].
4. New developments: not having good performance metrics that can automatically evaluate the perceived quality makes long and expensive subjective tests necessary [1].

Therefore, this situation proposal 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?

2.2 Scope of the work

Six no-reference metrics: FRIQUEE [9], QAWV [10], BRISQUE [11], NIQE [12], NSTSS [44], and TLVQM [14] will be evaluated in three publicly available databases: KoNVid-1K [15], LIVE Qualcomm [1] and LIVE Video Quality Challenge (VQC) [16] to extract data about the objective quality of authentically distorted videos in order to train a regressor for each no-reference video quality model to predict human scores of perceptual video quality.

3 Objectives

3.1 General Objective

To predict the video quality of three VQA databases KoNViD-1k [15], LIVE-Qualcomm [1] and LIVE Video Quality Challenge (VQC) [16], by applying six no-reference state-of-the-art VQA metrics FRIQUEE [9], QAWV [10], BRISQUE [11], NIQE [12], NSTSS [44] and TLVQM [14].

3.2 Specific Objectives

1. To extract video quality scores and features from three publicly available VQA datasets KoNViD-1k [15], LIVE-Qualcomm [1], and LIVE Video Quality Challenge (VQC) [16], according to six 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).

4 Justification

The results of the objective study that compares six no-reference metrics are particularly desirable in networked visual communication applications for the purpose of monitoring the quality of service (QoS) [17]. 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 [17]. Therefore, accurate and high-speed measurements of VQA perception can play an important role [17] and benefit from this in quality control and new developments such as:

1. Monitoring: the observation of video quality in the applications used by users, given the systematization of the measurements produced by the objective NR algorithms [18].
2. Quality control and management: when users pay for applications of a certain quality, it is necessary to predict the perceived quality in order to control the quality for which the user is paying [19].
3. New developments: on new advances to improve the ability of digital cameras, smartphones, computers, and tablets to acquire, display and play high-resolution images and videos [20].
4. Optimizing no-reference metrics: The no-reference FRIQUEE metric can take 5 hours to process a single video (with a frame size of 1280x720 pixels, depending on the tests made with the availables computers), so the reduction of the execution time can help to be used in applications or systems that require quality measurements at high speeds [9].

5 Theoretical framework

5.1 State of the art

5.1.1 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 [23].

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" [24]. Subjective methods are expensive, difficult, and impractical to perform in real-time applications [24], so it is necessary to use and develop objective and automatic methods, capable of reliably predicting perceived quality. Table 1 describes the VQA database to be used in this study.

5.1.2 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.

5.1.3 Authentic, 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 [26], [9].

Dataset characteristic	KoNViD-1k [15]	LIVE-Qualcomm [1]	LIVE-VQC [16]
Number of test videos	1200	208	585
Video resolution	960x540	1920x1080	320x240-1920x1080
Video frame rate	23-29 frames/sec	30 frames/sec	19-30 frames/sec (One sequence 120 frames/sec)
Video length	8 sec	15 sec	10 sec
Number of scenes	1200	54	585
Number of devices	>164	8	101
Test methodology	Crowdsourcing	Lab-based	Crowdsourcing
Number of test subjects	642 (min. 50 per video)	39	min. 200 per video
Rating scale	Absolute Category Rating (1-5)	Continuous 0-100	Continuous 0-100
Audio track included	Yes (some)	No	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.
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.
Remarks	Some contents in the database are clipped from the original.	Additional information collected concerning the dominating distortion type of each test sequence.	At the time of writing, not yet publicly available for download.

Table 1: Public consumer video quality databases compared: KONVID-1K, LIVE-QUALCOMM, and LIVE-VQC [14].

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 inex-

perienced 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, so there is no systematic way to synthetically add distortions to videos to accurately simulate authentically distorted videos [1].

5.1.4 Databases

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

KoNViD-1k [15] consists of a VQA database with 1200 public domain video sequences. The videos are encoded at three frame rates: 24, 25, and 30 Fps corresponding to 27%, 5%, and 68% of the videos in the database, respectively [15]. There are a total of 12 resolutions, but the largest percentage (85%) of the videos have a frame size of 1280x720 pixels, followed by 1920x1080 (9%), and most videos (97%) have an audio channel [15]. Due to the large number of videos, the subjective scores were obtained using the CrowdFlower platform at crowdflow [15].

LIVE-Qualcomm [1]: The database has a total of 208 videos. All videos were captured at a resolution of 1920x1080. They 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 [1].

LIVE Video Quality Challenge (VQC) [16]: 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. It contains a large number of subjective video quality scores through the crowdsourcing tool, with an average of 240 human opinions recorded per video [16].

5.1.5 Algorithms

Algorithms can be created that learn the human responses to distortion by training them in large databases of human opinion scores [20], for this project we used six NR image quality assessment (IQA) and VQA metrics:

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 so, it combines a large and diverse collection of statistical features of perceptually relevant real-world images through 4 layers [9].

These layers are called FRIQUEE-Luma, FRIQUEE-Chroma, FRIQUEE-LMS and FRIQUEE-ALL; FRIQUEE-Luma makes use of $a-f$ feature maps; FRIQUEE-Chroma uses g and h feature maps; FRIQUEE-LMS uses i and j feature maps; and 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 [9].

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. 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 [9]. 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 [9].

QAWV

Quality Assessment of In-the-Wild Videos (QAWV) is a NR VQA method that incorporates two eminent effects of the HVS in a deep neural network, these effects being content-dependent and temporary memory effects. Content feature maps are extracted through a neural network of image classification trained to analyze temporal memory effects and long-term dependencies (especially temporal hysteresis) using a subjectively inspired temporal grouping layer [10]. It should be noted that the human vision system (HVS) can be used for image compression (in which the highest frequencies are seen more precisely quantified) and motion estimation (using luminance and ignoring color). Also, HVS is used to simplify the behavior of complex systems [10].

BRISQUE

Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) is a blind/no NR-reference IQA evaluation metric that operates in the spatial domain and uses the scene statistics of locally normalized luminance coefficients to quantify possible losses of naturalness in the image due to the presence of distortions, resulting in a holistic measure of quality [11]. BRISQUE uses an SVR trained by a set of characteristics derived from the empirical distribution of luminances and locally normalized luminance products, taking into account a spatial natural scene statistical model [11].

However, BRISQUE does not require any transformation to another coordinate frame (DCT, wavelet, etc.), differing from the previously mentioned IQA approaches without reference. In addition, this metric does not calculate the characteristics of specific distortions such as ringing, blurring or blocking and has a low computational complexity, making it suitable for real-time applications [11].

NIQE

Natural Image Quality Evaluator (NIQE) is a completely blind image quality analyzer that uses only the measurable deviations from statistical regularities observed in natural images. However, all current general-purpose NR IQA algorithms require knowledge of anticipated distortions in the form of training examples and corresponding human opinion scores [12].

NIQE is based on the construction of a quality-aware collection of features in a natural scene statistics model in the spatial domain (NSS). These characteristics are derived from a corpus of undistorted natural images. The experimental results show that the new index offers a performance comparable to that of higher performance NR IQA models that require training in large databases of human opinions of distorted images [12]. NIQE, is an unsupervised technique as it is powered by NSS-based features, and does not require exposure to distorted images, which gives it even more generality.

NSTSS

No-Reference Video Quality Assessment Using Natural Spatiotemporal Scene Statistics (NSTSS) is based on a parameterized statistical model for the spatiotemporal statistics of mean subtracted

and contrast normalized (MSCN) coefficients of natural videos. Specifically, the authors propose an asymmetric generalized Gaussian distribution (AGGD) to model the statistics of MSCN coefficients of natural videos and their spatiotemporal Gabor bandpass filtered outputs. The authors then demonstrate that the AGGD model parameters serve as good representative features for distortion discrimination [44].

TLVQM

Two Level Video Quality Model (TLVQM) is a No-Reference Video Quality Model (NR-VQM) that is specifically designed to evaluate consumer video quality typically impaired by capture artifacts such as sensor noise, motion blur, and camera shake [14].

TLVQM is based on the idea of computing the most disturbing motion features of video hierarchically in two steps: low-complexity spatial-temporal features are computed from every second frame, and more complex spatial features are computed only from a subset of representative frames [14].

These features and the respective subjective quality scores are used to train two regression models, SVR and Random Forest Regression (RFR), in order to determine the MOS of objective quality prediction [14].

5.1.6 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 [27]-[30]. Initially, the relevant features of the videos to be studied are extracted in order to refine the existing quality measures [28], 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 [29], [30]. 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 [28], [10]. Inspired by this work [28], the features will be extracted from 1,993 videos (contained in the databases mentioned above) and then trained into a statistical regression model.

5.1.7 Regression

These features of the perceptually relevant videos, along with the corresponding real-value MOS of the training set, are used to train a SVR. 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 [9].

SVMs and regressors are 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 [9]. Although the databases to be used are large, they do not have the necessary robustness to motivate the use of deep learning methods.

5.1.8 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, this because it has 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 2 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 2. 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 3.

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

Equation 3. 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.

5.1.9 Data analysis measures

PLCC: Also known as the Linear Correlation of the Pearson Coefficient, it is a linear measure between two quantitative random variables (x,y) [33]. 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 [34]. Equation 4 is used to calculate the PLCC for a population.

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

Equation 4. 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 [33].
On the other hand, equation 5 is used to calculate the PLCC for a statistical sample

$$rxy = \frac{n \sum xiyi - \sum xi \sum yi}{\sqrt{n \sum xi^2 - (\sum xi)^2} \sqrt{n \sum yi^2 - (\sum yi)^2}}$$

Equation 5. 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 [33]:

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.

SROCC: Spearman's Rank-Order Correlation Coefficient, is a measure of correlation or interdependence between two random variables X and Y, such variables can be both continuous and discrete [35].

SROCC is denoted by the symbol rs (or the Greek letter p, pronounced rho) and it is calculated as equation 6.

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

Equation 6.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 [37].

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 [38]. Finally, the interpretation of SROCC is the same as that of PLCC.

RMSE: Also known as the Square Root of the Mean Square Error, 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 7:

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

Equation 7. 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.

5.1.10 Hypotheses considered

1. Of the blind VQA algorithms that have the longest runtime per frame and video are those that use a feature bag approach.
2. By incorporating time-space grouping strategies in blind VQA metrics to evaluate the LIVE-Qualcomm database, it improves the quality evaluation performance of these algorithms.
3. The highest PLCCs are obtained by those algorithms that incorporate neural networks and regressors.

5.1.11 Related work

Predicting the perceptual quality of an image or video that has been exposed to some kind of distortion has motivated several researchers to understand and make efforts to design systems capable of performing this task, which is why initially the documents [20] and [17] focus on showing how the current panorama of video transmission and reception is, as well as on indicating how the process of capture, compression and transmission can affect them by introducing distortions that impoverish the video quality for the end-users. During the development of these works the problems are announced and how to address them, carrying out different studies to understand the operation and seek strategies to solve them through the use of artificial intelligence (AI), this is where the idea of having a model that is capable of predicting the quality of video when they have been exposed to distortions is born.

In this order of ideas, the subjective studies arise as the first step towards the search of an AI system capable of predicting the VQA according to the type of distortion it has. To this end, in

[1] a database of videos taken under controlled conditions is made, with the aim of capturing videos by adding six different types of dominant distortions. Afterward, a subjective and an objective study is made analyzing the videos obtained in that database, for which four NR metrics were used to correlate the data obtained with the subjective study.

In the previous article, the comparison between the CVD-2014 and LIVE-Qualcomm databases is made. It is concluded that the latter is a potentially valuable tool that can be used to solve some of the limitations of the current VQA databases, in terms of content diversity, realism, and distortion variability. Likewise, CVD-2014 stands out for having multiple distortions in its unclassified videos.

That is why in the document [5], it is required to know how the videos are affected by authentic distortions. For this purpose, a new video database was designed that models a variety of complex distortions generated during the video capture process on portable capture devices. This article describes the content and features of the new database, and then goes on to conduct a subjective video quality assessment study using these data. Finally, several high-performance NR IQA/VQA algorithms were evaluated in the new database to understand how distortions in real-world capture challenge both human subjects and automatic predictive models of perceptual quality.

In the article [40], the authors try to answer how the quality of an image can be known if it has several criteria that can affect the perception of the quality of a final observer. To this end, it proposes the development and use of different methods capable of carrying out such an analysis. However, the following questions are raised regarding the validation of the results: which method or methods can be more efficient in the face of certain distortions, considering the computation times, how to describe a distortion from a limited sample of images in order to predict its results without prior knowledge?

Following the recommendations and possible future work set out at the end of each of the documents analyzed above, the idea of carrying out a study of the extraction of features from the videos was born. To this end, [41] proposes objective quality assessment methodologies focused on independent and relative segmentation, both for the individual object and for general assessment cases. These metrics expose the ability to estimate the quality of the segmentation according to what a human observer would do. For this purpose, the segmentation algorithm applied to the test sequences selected as representative of the application of the domain in question was carried out, then the object whose segmentation quality should be evaluated was selected and, finally, the evaluation of the quality of the segmentation was obtained.

On the other hand, the work [9] carried out a study of the perceptually relevant statistics of the natural settings of some images (databases), which were truly distorted in different color spaces and transformation domains. To extract the mapping bag from such images, demonstrating the competence of the features towards the improvement in the automatic prediction on the perceptive quality, to later train a regressor capable of analyzing and predicting the distortions that affect them.

In the quest to improve studies and results that can contribute more to a better understanding of the problem and obtain greater accuracy. In [10] it is proposed to evaluate the quality of the reference videos by integrating the effects in a neural network, to validate the performance of the metrics performed. Experiments were carried out in three publicly available VQA databases: KoNViD-1k , CVD2014, and LIVE-Qualcomm, respectively. These experimental results showed that the proposed method outperforms the results of the five most advanced methods, with a wide margin of 12.39%, 15.71%, 15.45% and 18.09% improvement in the overall performance compared to the second-best method, VBLIINDS, among the SROCC, KROCC, PLCC and RMSE methods, respectively.

Finally, most of the papers focus on the evaluation of NR image quality, which helps to detect the defects of the algorithms to be evaluated, allowing to determine the type of distortion in which they tend to make a better quality prediction, in addition to certain techniques that help to bring the results closer to those expected. According to the above, we will try to establish the connection between the subjective studies and the objective studies focused on the video analysis, generating a greater degree of difficulty in the evaluation of NR quality, using different NR metrics and databases, to later train a regressor with a good performance.

6 Research Methodology

6.1 Description

The proposed research is based on a quantitative, descriptive, correlative, and design approach. This is an analysis and data training investigation with the purpose of verifying and evaluating the behavior of NR VQA algorithms in videos with authentic distortions. The project's starting point is a bibliographic and theoretical review of subjective studies and NR video quality prediction metrics.

The methodology for developing the research is made up of 4 stages.

6.1.1 Stage 1: Extracting scores and features from NR VQA metrics

Specific objective:

To extract video quality scores and features from three VQA datasets KoNViD-1k [15], LIVE-Qualcomm [1], and LIVE Video Quality Challenge (VQC) [16], according to six no-reference video quality models.

Tasks:

1. To identify the programming language in which the NR VQA metric operates, as well as its operating system (Ubuntu, Windows).
2. To identify and install video quality prediction and image analysis libraries.
3. To set the code to perform frame-by-frame extraction of each video from the databases and processes it with the main function of the NR VQA metric.
4. To modify the code so that it stores the video quality scores of each frame and then perform the arithmetic average of this to obtain the overall score of the video.
5. To modify the code to store the video quality features of each frame.
6. To study the theoretical foundation of the technique called average pooling and then incorporate it into the Matlab analysis tool.
7. To study which are the best spatial-temporal grouping techniques.
8. To understand the theoretical foundation of the VQPooling technique and then implement it in Matlab.

6.1.2 Stage 2: Train regressor

Specific objective:

To train a regressor for each no-reference video quality model to predict human scores of perceptual video quality.

Tasks:

1. To identify the type of regressors to be used.
2. To do a theoretical analysis of the regressor to use and perform a bibliographic review of previous studies where regressors applied to the NR VQA metrics were utilized.

3. To identify the programming language in which the regressor will be trained.
4. To execute the regressor taking into account the execution time in the reading of the video quality prediction features and its analysis.

6.1.3 Stage 3: Performance analysis of NR VQA metrics

Specific objective:

To evaluate the performance of no-reference VQA metrics based on Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Correlation Coefficient (SROCC) and Root Mean Square Error (RMSE).

Tasks:

1. To set the functions to calculate the program execution time.
2. To identify the functions or lines of code that take the longest time to process.
3. To study the behavior of the functions and statistical methods used to obtain the video quality prediction score.
4. To study the theoretical foundation of PLCC, SROCC, and RMSE and then carry out its implementation in Matlab.
5. To analyze the results produced by the PLCC, SROCC, and RMSE.

6.1.4 Stage 4: Writing and reporting results.

The methodology should include a stage for writing and communicating the results of the analysis and research. Such writing must be done in the IEEE format with the structure required for a publication.

Figure 1 shows the flow chart, which highlights the main processes of each stage.

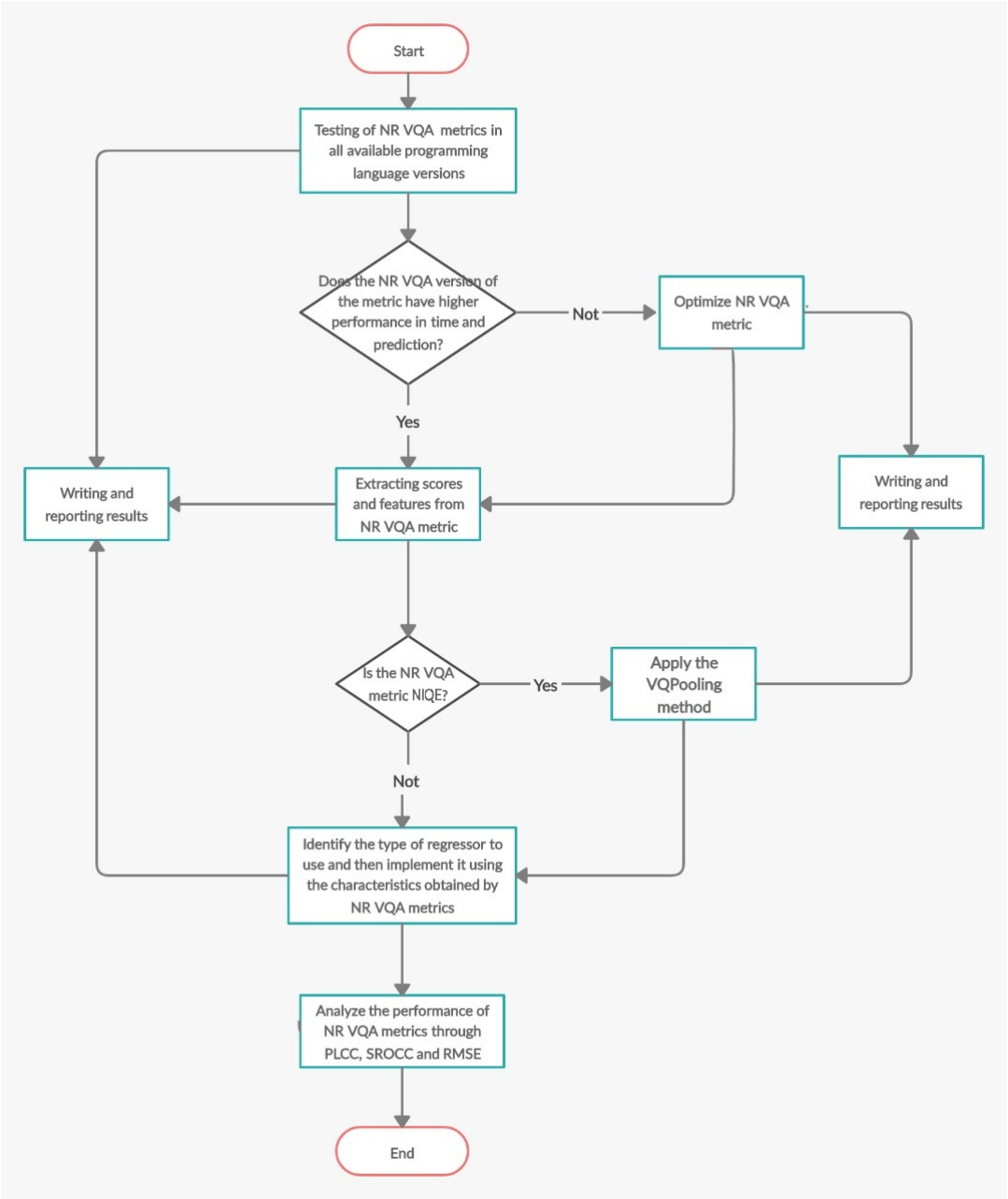

Figure 1. Flowchart of the methodology.

6.2 Tests

The libraries required to run the NR FRIQUEE [9] algorithm were installed. Once an image was processed, the necessary modifications were made in order to analyze each of the frames of a video from the LIVE Video Quality Challenge (VQC) database [16] extracting the scores and features vectors of each frame. After this, an analysis was performed with the objective of reducing the execution time of each video, obtaining a 34% improvement in processing time.

6.3 Expected Result

NR metrics are expected to correlate well (correlation coefficients close to 1) with human assessment of authentically distorted videos. In addition, the trained regressor is expected to predict the human assessment of video quality with authentic distortions better than the NR metrics studied.

7 Resources

7.1 Human

7.1.1 Director

Hernán Darío Benítez Restrepo
Received his undergraduate degree in Electronic Engineering, and his Dr. Eng. degree in Electronic Engineering from Pontificia Universidad Javeriana Sede Cali and Universidad del Valle, in 2002 and 2008, respectively.Dr. Benitez has been an IEEE Senior member since 2014 and Chair of Colombia's IEEE Signal Processing Chapter since 2012. He is a member of the scientific editorial board of the Quantitative Infrared Thermography Journal since 2014. He is the recipient of a Fulbright Visiting Researcher scholarship to carry out research on video quality assessment in the Laboratory of Image and Video Engineering (LIVE) at the University of Texas at Austin in 2019. His main research interests encompass image and video quality evaluation, infrared vision, and pattern recognition [36].

7.1.2 Co-director

Roger Gomez Nieto
Received his bachelor's degree in Electronic Engineering from the University of Quindio (2014), his Master's degree in Electrical Engineering from the Technological University of Pereira (2016), and is currently a Doctorate candidate at the Pontificia Universidad Javeriana Sede Cali. He did a six-month internship at the Wireless Networks and Communications Group at UT Austin, working with Professor Alan Bovik on Video Quality Assessment. His main research interests span deep learning and computer vision. His current research is focused on the design and test video object tracking algorithms explicitly robust in terms of performance with respect to in-capture and post-capture distortions [37].

7.1.3 Research group of the faculty that supports it.

Research Group: Automatic and Robotics Group (GAR)
Faculty of Engineering Pontificia Universidad Javeriana Cali

7.2 Economical

7.2.1 General resource's budget required

The project is expected to cost a total of 38.752.786 Colombian pesos or 9.900 US dollars..
Details of this amount are shown in Table 2.

Ruby	Unit value	Quantity(US-COP)	Quantity(US-COP)
MatLab License (Standard)	3	940 - 3.713.159	2.820 - 11.139.477
Adviser (40 hours)	2	607.57 - 2.400.000	1.215 - 4.800.000
Student (600 hours)	2	759.46 - 3.000.000	1.518 - 6.000.000
Image Processing Toolbox	3	6 - 23.701	18 - 71.103
Optimization Toolbox	3	6 - 23.701	18 - 71.103
Statistics and Machine Learning Toolbox	3	6 - 23.701	18 - 71.103
Server	1	3083,57 - 12.000.000	3083,57 - 12.000.000
Student computer 1	1	565.01 - 2.200.00	565.01 - 2.200.00
Student computer 2	1	607.57 - 2.400.000	607.57 - 2.400.000
Total			9.900 - 38.752.786

Table 2: Budget

Note: The prices were taken on April 30, 2020.

8 Schedule

The project schedule is described in Table 3.

		Months					
No	Activity	01	02	03	04	05	06
	Theoretical framework	x	x	x	x	x	
01	To identify the programming language in which the VQA NR metric operates, as well as its operating system (Ubuntu, Windows).	x					
	To identify and install video quality prediction and image analysis libraries.	x	x	x			
	To set the code to perform frame-by-frame extraction of each video from the databases and then be entered in the main function of the VQA NR metric.	x	x	x			
	To modify the code so that it stores the video quality scores of each frame and then perform the arithmetic average of this to obtain the overall score of the video.	x	x	x			
	To modify the code to store the video quality features of each frame.	x	x	x			
	To study the theoretical foundation of the technique called average pooling and then incorporate it into the Matlab analysis tool.			x	x	x	
	To study which with what are the best spatial-temporal grouping techniques.			x	x	x	
	To understand the theoretical foundation of the VQPooling technique and then carry out it's programming in Matlab.			x	x	x	
02	To identify the type of regressors to be used.		x	x	x	x	
	To do a theoretical analysis of the regressor to use and perform a bibliographic review of previous studies where regressors applied to the VQA NR metrics were performed.			x	x	x	x
	To identify the programming language in which the regressor will be performed.			x	x	x	x
	To perform the regressor taking into account the execution time in the reading of the video quality prediction features and its analysis.				x	x	x
03	To add the functions to calculate the program execution time.	x					
	To identify the functions or lines of code that take the longest to process.		x				
	To study the behavior of the functions and statistical methods that you use to obtain the video quality prediction score.		x	x	x		
	To study the theoretical foundation of PLCC, SROCC and RMSE and then carry out its implementation in Matlab.					x	x
	To analyze the results produced by the PLCC,SROCC and RMSE.					x	x
04	Writing and reporting results.	x		x	x		x

Table 3. Schedule

9 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] "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].

[3] "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].

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

[5] 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.

[6] N. Montard and P. Bretillon, "Objective quality monitoring issues in digital broadcasting networks," in IEEE Transactions on Broadcasting, vol. 51, no. 3, pp. 269-275, Sept. 2005, doi: 10.1109/TBC.2005.851700.

[7] Bretillon, P., Baina, J., Jourlin, M., Goudezeune, G. (1999, November). Method for image quality monitoring on digital television networks. In Multimedia Systems and Applications II (Vol. 3845, pp. 298-306). International Society for Optics and Photonics.

[8] M. H. Pinson, L. Janowski and Z. Papir, "Video Quality Assessment: Subjective testing of entertainment scenes," in IEEE Signal Processing Magazine, vol. 32, no. 1, pp. 101-114, Jan. 2015, doi: 10.1109/MSP.2013.2292535.

[9] 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.

[10] D. Li, T. Jiang, M. Jiang, "Quality Assessment of In-the-Wild Videos,", 2019.

[11] 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.

[12] A. Mittal, R. Soundararajan and A. C. Bovik, "Making a "Completely Blind" Image Quality Analyzer," in IEEE Signal Processing Letters, vol. 20, no. 3, pp. 209-212, March 2013.

[13] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," 2015 IEEE International Conference on Computer Vision (ICCV), 2015.

[14] J. Korhonen, "Two-Level Approach for No-Reference Consumer Video Quality Assessment," in IEEE Transactions on Image Processing, vol. 28, no. 12, pp. 5923-5938, Dec. 2019, doi: 10.1109/TIP.2019.2923051.

[15] 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.

[16] 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.

[17] 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.

[18] Le Callet, P., Viard-Gaudin, C., Péchard, S., Caillault, E. (2006). No reference and reduced reference video quality metrics for end to end QoS monitoring. IEICE transactions on communications, 89(2), 289-296.

[19] Perceptual Quality Measurement and Control: Definition, Application and Performance AR Prasad, R Esmailzadeh, S Winkler, T Ihara, B 4th International Symposium on Wireless Personal Multimedia Communications, Aalborg, Denmark, 2001

[20] A. C. Bovik, "Automatic Prediction of Perceptual Image and Video Quality," in Proceedings

of the IEEE, vol. 101, no. 9, pp. 2008-2024, Sept. 2013.
[21] Recommendation ITU-R BT.500-11 Methodology for the subjective assessment of the quality of televisión pictures, 06/2002
[22] Recommendation ITU-T P.910 Subjective video quality assessment methods for multimedia applications, 09/1999
[23] 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.
[24] 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.
[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] Ernestasia Siahaan,Alan Hanjalic,and Judith A Redi.2018. Semantic-aware blind image quality assessment. SPIC 60(2018),237–252.
[30] 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.
[31] Tu, Z., Chen, C. J., Chen, L. H., Birkbeck, N., Adsumilli, B., Bovik, A. C. (2020). A Comparative Evaluation of Temporal Pooling Methods for Blind Video Quality Assessment. arXiv preprint arXiv:2002.10651.
[32] 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.
[33] “SPSS Tutorials: Pearson Correlation,” LibGuides. [Online]. Available: libguides.library.kent.[Accessed: 30-Apr-2020].
[34] “Basic Concepts of Correlation | Real Statistics Using Excel.” [Online]. Available: http://www.real-statistics.com/correlation/basic-concepts-correlation/. [Accessed: 30-Apr-2020].
[35] A. Lehman, JMP for basic univariate and multivariate statistics. Cary, NC: SAS Press, 2005, p. 123.
[36] Myers and A. Well, Research design and statistical analysis, 2nd ed. Mahwah, N.J.: Lawrence Erlbaum Associates, 2003, p. 508.
[37] “The Concise Encyclopedia of Statistics,” 2008.
[38] A. Al Jaber and H. Elayyan, Toward quality assurance and excellence in higher education. River Publishers, 2018, p. 284.
[39] E. Lehmann and G. Casella, Theory of point estimation, 2nd ed. New York: Springer, 1998.
[40] Z. Wang, “Objective Image Quality Assessment: Facing The Real-World Challenges,” Electronic Imaging, vol. 2016, no. 13, pp. 1–6, 2016.
[41] P. L. Correia and F. Pereira, ”Objective evaluation of video segmentation quality,” in IEEE Transactions on Image Processing, vol. 12, no. 2, pp. 186-200, Feb. 2003.
[42] 2020. [Online]. Available: https://co.linkedin.com/in/hernan-dario-benitez-restrepo-a828093. [Accessed: 30- Apr- 2020].
[43] 2020. [Online]. Available: https://www.linkedin.com/in/rogergom/. [Accessed: 30- Apr-2020]. Li, Y., amp; Hu, X. (2017).
[44] Li, Y., amp; Hu, X. (2017). No-Reference Stereoscopic Image Quality Assessment Using Natural Scene Statistics. 2017 2nd International Conference on Multimedia and Image Process-

ing (ICMIP). doi:10.1109/icmip.2017.61.