IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 14 February 2024, accepted 27 February 2024, date of publication 4 March 2024, date of current version 8 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3373193

RESEARCH ARTICLE

Efficient Computational Cost Saving in Video Processing for QoE Estimation

ÁLVARO LLORENTE [1], **JAVIER GUINEA PÉREZ** [1,2], **JUAN ANTONIO RODRIGO** [3], **DAVID JIMÉNEZ** [4], **(Member, IEEE), AND JOSÉ MANUEL MENÉNDEZ** [1], **(Senior Member, IEEE)**

[1]Señales, Sistemas y Radiocomunicaciones, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, 28040 Madrid, Spain
[2]Video-Mos, 28001 Madrid, Spain
[3]Sistemas Informáticos, Escuela Técnica Superior de Ingeniería de Sistemas Informáticos, Universidad Politécnica de Madrid, 28031 Madrid, Spain
[4]Electrónica Física, Ingeniería Eléctrica y Física Aplicada, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, 28040 Madrid, Spain

Corresponding author: Álvaro Llorente (alg@gatv.ssr.upm.es)

ABSTRACT No-Reference video quality assessment has become a trending and challenging hot topic in estimating perceived quality in audiovisual content. In this paper, we present a proposal to considerably reduce the computational cost of video processing without losing accuracy in QoE estimation. Tests have been performed using the Video-MOS SaaS solution, a hybrid NR-VQA solution based on perceptible video distortions and a machine learning approach. After exploring the spatial and temporal redundancy present in a video sequence, the final approach combines video metric feature extraction in both high and low video resolution, together with a specific frame selection based on a uniform temporal sampling and frame type at the video coding level. An extensive validation with more than 144 hours of audiovisual content from six of the most important HD channels of DTT in Spain demonstrates the validity of the approach, ensuring real-time application on the test device, with computational cost savings of 94.96% and an obtained MOS error of 0.1144, in more than 174000 3-second measurements.

INDEX TERMS Computational cost, feature extraction, I frames, machine learning, mean opinion score (MOS), no-reference, video quality assessment, perceived quality, quality of experience (QoE), video processing.

I. INTRODUCTION

Audiovisual content traffic has grown considerably over the last few years. The massive use of social networks, improvements in Internet speed and connectivity, and new audiovisual consumption habits have led to a huge boom in media applications and services: video surveillance, virtual reality, augmented reality, Internet Protocol TV (IPTV), Video-on-Demand (VoD) and gaming. Video has become an increasingly important part of global Internet traffic. IP video traffic has been estimated to be 75% of all IP traffic by 2017 and 82% by 2022 [1]. Video streaming services such as YouTube, Netflix, Facebook Video, and TikTok account for a large part of the IP video traffic [2].

A. THE IMPORTANCE OF QoE ESTIMATION

The success of audiovisual content or a media application is directly related to the end-user satisfaction. Measuring the perceived quality by end-users has become one of the most important goals for broadcasters and content providers. There are many processing stages from the audiovisual content acquisition to its consumption. All of them produce distortions that can affect the final perceived quality. Contrast or color issues due to the nature of the scene, blurring, freezing, block effect, bitrate loss, packet loss or latency could be some of the typical distortions produced in the audiovisual chain.

Although spatial consistency (such as the realism of object shapes, color, and textures) or temporal consistency (such as the movement of objects) are the main factors in perceived quality [3], subjectivity is not easy to measure. The typical process used to assess the perceived quality is known as QoE (Quality of Experience) estimation. Using the ITU's (International Telecommunication Union) definition, QoE is "the overall acceptability of an application or service, as perceived subjectively by the end user" [4]. This measure considers the type of content, signal degradations, expectations, experiences, and user perceptions related to the

The associate editor coordinating the review of this manuscript and approving it for publication was Yun Zhang.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

Human Visual System (HVS) and Human Auditory System (HAS), network conditions, and device capabilities. Many issues are still open in QoE field due to the multiple human, system, and content influencing factors [5].

B. IMAGE/VIDEO QUALITY ASSESSMENT

Image Quality Assessment (IQA) and Video Quality Assessment (VQA) have been studied extensively over the last decade to measure the QoE. VQA can be divided into two categories: subjective quality assessment and objective quality assessment.

Subjective quality assessment is the most reliable way to assess the perceived quality since videos are aimed at end-users. Subjective quality is measured by asking a human subject to indicate the quality of an image or video, typically using a numerical scale, such as MOS (Mean Opinion Score) scale, with five possible values (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). Statistical significance of the MOS value must be guaranteed. Several assessment methodologies have already been standardized by the ITU in ITU-T Recommendation P.910 [6] and ITU-R Recommendation BT.500 [7]. These methodologies describe in detail how subjective video quality experiments should be set up and conducted. Due to the strictness of the methodologies, subjective assessments are time-consuming, expensive, and impractical for real-time applications.

Objective quality assessment predicts the perceived video quality scores automatically with computational VQA models that simulate the HVS and human perception. Objective VQA performance has already been widely investigated by the Video Quality Experts Group (VQEG). These assessments can be categorized into three categories based on the availability of the original video: Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) or Blind VQA (BVQA). Another criterion to categorize objective VQA is the type of information extracted from the video sequence: pixel-based, bitstream-based, parametric-based, or hybrid, being a combination of all of them.

C. CHALLENGES IN QoE ESTIMATION

The development of an objective video metric that accurately estimates the perceived video quality is still challenging nowadays. Not only because of the task of finding an algorithm whose quality prediction is in good agreement with subjective scores from real human observers [8], but also because of the emergence of new types of content and applications that are clearly differentiated from traditional audiovisual content, and require the design of specific video metrics: User-Generated content (UGC) [9], High Dynamic Range (HDR) audiovisual content [10], [11], omnidirectional videos [12], [13], [14], [15], videogames [16], and artificial and enhanced videos [17], [18], [19].

Another important challenge is the progress with the new video formats. Video resolutions are continuously increasing to provide more realistic and immersive experiences. Following the success of High Definition (HD) video services, the Ultra High Definition (UHD) format [20] is now a reality and is considered the future standard for video applications. Popular video streaming platforms such as YouTube, Netflix, or Amazon already support 4K UHD resolution videos.

The study of subjective and objective VQA is necessary for these new video formats [21], [22]. There is a major technological challenge in the design of objective video metrics for 4K and 8K video resolution with high frame rates. The spatial resolution of 4K UHD content [23] is four times the Full HD resolution [24]. And there is sixteen times more information between 8K UHD resolution [23] and Full HD resolution.

D. VIDEO-MOS SaaS SOLUTION

The motivation for this work and this study arises from these challenges in the objective video metrics field. Video-MOS SaaS (Software as a Service) solution is a video content quality monitoring commercialized by the European company Video-MOS [25]. The solution is a hybrid NR-VQA system based on perceptible distortions and a machine learning-based approach. Thanks to its real-time operation and its advanced Artificial Intelligence technology, this SaaS solution can perform a complete QoE monitoring in terms of MOS value estimation, specific distortion detection, and impact generated on the end-user [26], [27].

Video-MOS SaaS solution is protected at Registro Territorial de la Propiedad Intelectual de la Comunidad de Madrid (*Territorial Registry of Intellectual Property Right of the Community of Madrid region*), with the registration of four software modules: M-002018/2023, M-002033/2023, M-002037/2023 and M-002039/2023. It is also under patent application.

E. OBJECTIVE AND CONTRIBUTIONS

The main objective of this work is to reduce the computational cost of a hybrid NR-VQA assessment tool, maintaining correct monitoring performance and accuracy in QoE estimation. For the VQA measurements, the Video-MOS SaaS solution has been used. Savings in computational costs have multiple benefits such as real-time processing of UHD content or processing a greater number of contents on the same device, thus allowing for significant financial savings, reducing infrastructure costs (space and hardware), lower energy consumption, flexibility, and improved scalability. Although there are different strategies to reduce computational costs in video processing using specific hardware (eg. Graphics Processing Units (GPUs)) or parallelization techniques and distributed processing, the work will focus on exploring the spatial and temporal redundancy that characterizes video sequences. Different subjective and objective studies have analyzed the impact of spatial and temporal subsampling [28], [29], [30], [31], [32], [33], [34], [35], [36]. This study is open to the application of these same computational cost reduction techniques to other IQA/VQA measurement proposals.

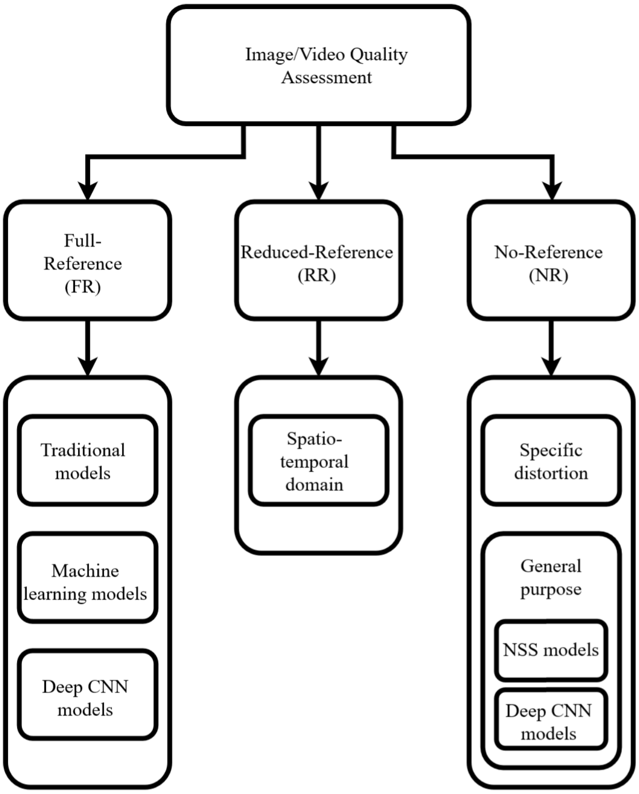

FIGURE 1. Overview of VQA methods.

II. RELATED WORK

Researchers in IQA and VQA fields have been working to understand how distortions introduce a degradation in the audiovisual signal and how it impacts signal statistics and perceived quality. There has been a steady evolution from traditional models to learning-based and deep-learning techniques. Fig. 1 provides an overview of the literature collected in this section. With this advancement, the feature extraction has improved to achieve better prediction performance when compared with classical approaches. Deep VQA modeling is a field that still needs a lot of research. There is a major limitation because of the lack of reliable large and diverse training databases and ineffective training methods [37]. Small databases are insufficient for training models with relatively high network capacity and for detecting multiple specific video distortions simultaneously. Additionally, these models trend to be overfitted.

A. FULL-REFERENCE QUALITY ASSESSMENT

FR IQA/VQA models require the presence of a reference signal to predict the quality of the distorted signal. The simplest monitoring approach is to compare the original with the received video and measure the differences. The degradation or loss of quality is calculated based on the measured deviation. However, the non-availability of the reference limits the use of FR metrics in many applications.

Traditional FR IQA models measure frame-by-frame deviation metrics such as MSE and PSNR [38], [39]. Both are efficient but often offer poor correlation with subjective perception. Other FR IQA models achieve better correlations with subjective scores and visual perception: PSNR based on HVS [40], SSIM [41], MS-SSIM [42], VSNR [43], MAD [44], VIF [45], FSIM [46] and FMSE [47]. However, the demonstration over time that motion information plays an important role in the visual perception influencing the perceived quality (HVS is more sensitive to distortions on moving objects because the movement automatically attracts attention), led to the appearance of spatio-temporal VQA models such as MOVIE [48], VIS3 [49], ST-MAD [50], PVM [51], FLOSIM-FR [52] with optical flow information, or FAST [53] with salient trajectories information.

More sophisticated FR approaches make use of machine learning techniques such as VMAF [54], a solution developed by Netflix that proposes the use of multiple VQA features with learning-based regression, ST-GREED [55] with a support vector regressor, or [56] with a random forest regression algorithm used to map multiple features (texture, saliency, spatial activity, and temporal activity) into a subjective score.

Latest FR approaches use deep convolutional neural networks (CNN) such as DeepVQUE [57], DeepVQA [58], C3DVQA [59], [60], [61], DISTS [62], DeepQA [63], and CONTRIQUE-FR [64]. All of them have demonstrated the potential to compete with traditional metrics, but the lack of subjective databases make them limited models for different types of content and for specific distortions.

B. REDUCED-REFERENCE QUALITY ASSESSMENT

RR IQA/VQA models require only partial information about the reference signal to predict the quality. These models also exploit the spatio-temporal information, extracting information in the spatial domain, temporal domain or combining both domains: RRED, TRRED, ST-RRED [65] and SPEED QA [66].

C. NO-REFERENCE QUALITY ASSESSMENT

NR or Blind IQA/VQA models have greater potential and wider application than the FR and RR models by being able to predict the quality without the need for reference signal information. Existing BVQA models are often designed based on two approaches: specific distortion or general purpose.

Specific distortion approaches focus on estimating perceived quality in contents that have a particular type of distortion such as artifacts [67], block effect distortion [68], [69], blur and noise [70], [71], [72], ringing [73], [74] or banding [75]. However, these models cannot be extended to real-world videos, which contain many types of combined spatial and temporal distortion.

General purpose approaches are based on (multi-)feature extraction and learning-based techniques, training a set of generic quality-aware features combined to conduct the quality predictions. The possibility to extract relevant perceptual features combined with the use of powerful regression models make general-purpose methods much more versatile and generalizable than specific distortion approaches. In general, learning-based approaches either use regression or classification for the perceived quality estimation: regression is commonly used for MOS value estimation whereas classification is typically used for predicting error visibility by means of binary decision.

Most popular BVQA algorithms employ perceptually relevant low-level characteristics such as natural statistical features of the images based on Natural Scene Statistics (NSS) models [76]. NSS models are based on the idea that the distortion in a natural image can change the natural statistical features of the scene, making the image unnatural. Successful NSS general-purpose models have been proposed exploring the structural information in the DCT (Discrete Cosine Transform) domain (BLIINDS [77], BLIINDS-II [78]), spatial domain (NIQE [79], BRISQUE [80]), wavelet domain (BIQI [81], DIIVINE [82]) and gradient-domain (GM-LOG [83], [84], HIGRADE [85]). FRIQUEE [86] achieves good performance predicting the perceptual quality of images corrupted by a combination of multiple authentic distortions. CORNIA [87] is efficient, effective, and computationally fast. VIDEVAL [88] focuses on spatial distortions selecting a combination of simple distortion-aware statistical video features, NSS statistics, and well-defined visual impairment features.

VBLIINDS [89] was one of the first models to explore the use of spatiotemporal NSS in the time-differenced domain, computing motion coherence and global motion features with expensive motion estimation operations. VIIDEO [90] and 3D-DCT NR-VQA [91] exploit a greater variety of spatio-temporal statistical regularities to predict and quantify the quality of distorted videos. STFC [92] model also extracts spatiotemporal statistics and achieves good performance with authentic distortions by being designed using authentic distorted videos. ChipQA [93] model is based on a quality-aware feature (space-time chips) in localized spatiotemporal cuts in directions determined by the local motion flow.

TLVQM [94] model captures artifacts such as camera shakiness, overexposure, underexposure, and sensor noise in UGC videos. This model uses spatio-temporal feature extraction making use of a mechanism for selecting the frames used for computing different types of features: low complexity features from full video and high complexity features from representative video frames. This mechanism considerably reduces the computational cost of TLVQM model.

Recently, several deep CNN-based BVQA models have been proposed: PATCH VQ [95], MLSP VQA [96], GSTVQA [97], RankDVQA [37] and DEEPSTQ [98].

CNN-TLVQM [99] improves the TLVQM model by replacing the spatial high-complexity features with deep features. VSFA [100] proposes the integration of the content-dependency effect and the temporal-memory effect into deep neural networks (DNN), and MDTVSFA [101] is an enhanced version of the VSFA.

DisCoVQA [102] method aims to model both temporal distortions and content-related temporal quality attention via transformer-based architecture. COINVQ [103] model proposes a DNN-based framework to thoroughly analyze the importance of content, technical quality, and compression level in perceptual quality for UGC videos. Li et al. [104] propose a transfer learning method for in-the-wild scenarios to leverage knowledge from spatial appearance and temporal motion.

V-MEON [105], STFEE [106], and SACONVA [107] use a 3D CNN for spatio-temporal feature extraction and evaluation. RAPIQUE [108] model exploits and combines efficiently spatial and temporal scene statistics as well as deep spatial features of natural videos, achieving good performance.

The main limitation of all these models lays on the restricted size of datasets available for training neural networks. In any case, they would all be able to benefit from the proposals put forward in this work as well.

In addition, NR-VQA models are often computationally complex and impractical for many real-life applications when evaluating videos of HD and beyond resolutions. Recent work focuses on efficiently modeling the spatial and temporal information of a video sequence, improving the performance of VQA models, with the goal of reducing computational cost and hardware requirements without compromising the accuracy of video quality prediction.

In video comprehension tasks pursuing the trade-off between effectiveness and efficiency, some researches tried to reduce the number of input frames by sparse sampling, taking into account that there is a lot of redundant information in consecutive frames. In this work [109], the proposed method exploits a novel sampling module capable of selecting a predetermined number of frames from the whole video sequence. With a substantially lower computational cost, the algorithm removes temporal redundancy by selecting a set of representative frames and achieves promising performance. In [110], different frame sampling strategies were designed. The findings of this study show that sparsely sampled video frames can obtain a competitive performance against using all video frames for quality estimation.

Apart from exploiting the temporal redundancy of the video, other proposals also take advantage of the spatial redundancy of the image, using regions of interest for feature extraction or downsampled images. The NR-VQA model proposed in [111] uses a systematic sampling of the three spatiotemporal planes, and the one proposed in [112] combines frame sampling strategy with a multi-resolution patch sampling mechanism to maintain the high-resolution quality information. The work done in [113] integrates the fusion of temporal statistics of local and global image features. Zoom-VQA [114] proposes an architecture to perceive spatiotemporal features at different levels, efficiently capturing both local and global information in regions of interest and in the whole frame. FAST-VQA [115] is based on a video sampling scheme that preserves quality by using fragments of the image rather than considering naive sampling approaches such as resizing and cropping. Finally, DOVER [116] proposes two independent quality evaluators that use spatial downsampling and temporal sampling of sparse frames to learn semantic and contextual information, and sampled raw resolution patches to form fragments similar to those introduced in FAST-VQA.

The strategies applied in these recent models bring benefits and higher efficiency to state-of-the-art NR-VQA methods. This is a good starting point to focus our work on reducing the complexity of our hybrid NR-VQA assessment tool.

III. METHODOLOGY

Video is a sequence of consecutive frames usually very similar to each other (temporal redundancy). Within a frame, a pixel also maintains a similarity with neighboring pixels (spatial redundancy). In the same way that video encoders use techniques based on spatial and temporal redundancy to compress and reduce the amount of information in a video signal, the proposed approaches to save computational costs in quality estimation will also focus on exploring these two types of redundancies. Processing smaller images and/or processing a reduced number of images in a video sequence can considerably decrease the computational cost.

A. TESTING TOOL

The measure used for quality estimation in this study is the estimated MOS calculated using the hybrid NR-VQA estimator from Video-MOS. Two main advantages made this metric suitable for our objective: Firstly the solution uses statistical descriptors of the video feed of both spatial and temporal information, allowing redundancies in both domains to be exploited. The other benefit of using the MOS estimation from this particular software solution is that Video-MOS has an agreement of collaboration with Universidad Politécnica de Madrid as a research chair [117] allowing for full access to the tool and on-demand changes to its functioning for research purposes.

The solution used for this study is the Video-MOS development tool. This tool includes all the functionalities of the commercial version and offers the same results in terms of feature extraction, MOS value estimation, and specific distortion detection. The main difference between both tools is that the development tool is built in Python instead of C++. This means that the development tool is much less computationally efficient than the commercial version, but its ease of making quick changes when proposing different approaches makes it the ideal tool for the study intended in this work. And, of course, any improvements made to the development tool will make it possible to improve the commercial version as well.

Feature extraction in the Video-MOS SaaS solution consists of a set of features that spatially and temporally characterize a set of frames of a video sequence. The solution uses a non-linear regression model based on artificial intelligence to process a set of parameters from the hybrid analysis of the video signal, with video metadata, NR video metrics and specific video distortion detection. The learning-based techniques estimate the numerical value of the perceived video quality within the range of the MOS scale according to the ITU-R BT.500 [7]. Table 1 lists some of the parameters used for the quality estimation.

TABLE 1. Features used for the MOS estimation. Combination of video metadata, NR video metrics and specific video distortions.

Type	Parameters
Video metadata	Resolution Frame rate Scan type Video codec Bitrate Bit depth Chroma subsampling Color space
NR video metrics	Spatial Information Temporal Information Blurring Brightness Contrast Ringing Blockloss Blocking
Specific video distortions	Block effect Artifacts Frame loss Content loss Signal loss Bright frames Dark frames Freezing Contrast High/Low Saturation High/Low Overexposure Underexposure

TABLE 2. Main characteristics of the HD format in DTT in Spain.

Parameter	Value
Resolution	1920x1080
Aspect ratio	16:9
Frame Rate	25 frames-per-second
Scan Type	Interlace
Chroma subsampling	YCrCb 4:2:0, 8 bits
Colour Space	ITU-R BT.709
Video encoding	H.264/MPEG-4 AVC

The quality estimation is done in user-defined measurement intervals. However, for testing purposes, an interval of 3-second measurements has been established. The tool estimates the MOS value every three seconds using the features extracted from the set of frames belonging to that time interval of the video sequence.

B. TEST SEQUENCES

The set of test sequences is composed of 1123 3-second measurements in HD format used on DTT (Digital Terrestrial Television) in Spain. Table 2 summarizes the main characteristics of this format. The test set also includes more than 84000 individual images corresponding to the

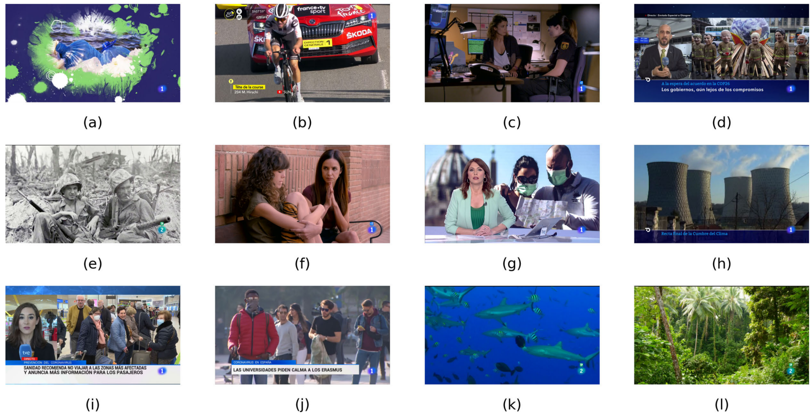

FIGURE 2. **Test sequences. Screenshots RTVE contents [119], [120]. Type of content: synthetic content and graphics (a), old black and white content (e), nature documentaries (k, l), indoor and outdoor news (d, g, h, i, j), sports (b), series and movies (c, f).**

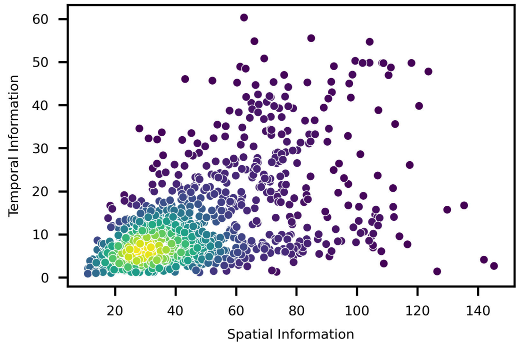

FIGURE 3. **Test sequences. SI-TI diagram.**

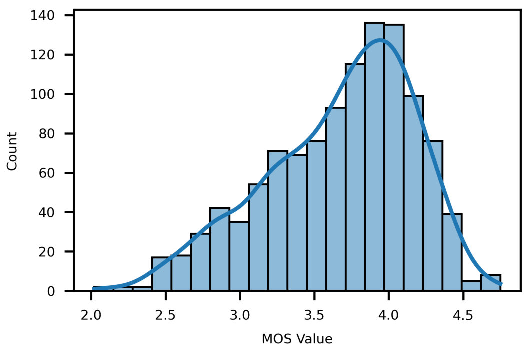

FIGURE 4. **Test sequences. MOS value histogram.**

1123 measurements. A frame rate of 25 frames per second means 75 frames in a 3-second video measurement.

Contents have been obtained directly from the DTT broadcasting using professional equipment, tuning two DTT multiplex (RGE1 and RGE2) [118] where the public broadcaster RTVE (Radiotelevisión Española) [119] offers its television channels in Spain. The sequences contain a wide variety of content, including pieces of news, sports, musicals, documentaries, movies, and series. RTVE and Universidad Politécnica de Madrid signed an agreement in the form of a University Chair in 2015 [120]. The contents used in this test have the explicit permission of RTVE for R&D activities within this project.

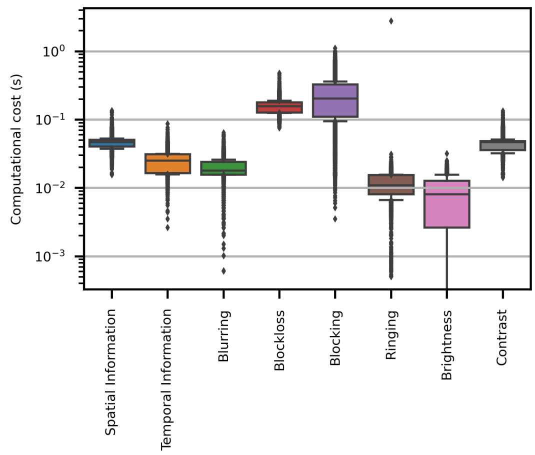

FIGURE 5. **Feature extraction time per video metric. Boxplot representation in logarithmic scale.**

The test set has a great diversity in the type of content it includes: synthetic content with the presence of graphics, old black and white content, documentaries, indoor and outdoor news, sports, series, and movies. Fig. 2 shows some screenshots of the test video sequences.

The diversity of the sequences is also manifested in the wide range of SI (Spatial Information) and TI (Temporal Information). Fig. 3 depicts the SI-TI diagram of all sequences. SI and TI values are calculated according to the expressions in ITU-T Recommendation P.910 [6], edition 4.0 (November 2021). In terms of MOS value of the 1123 3-second measurements obtained directly by the Video-MOS SaaS tool in normal processing mode, there is also a variation in the perceived video quality. Fig. 4 shows the histogram of the MOS values.

Most of the sequences have a MOS value higher than 3 (Fair on the MOS scale). The mean MOS value among all the sequences is 3.67, the maximum value is 4.75 and the minimum value is 2.02. There is a set of 131 measurements (11.67%) with a MOS value of less than 3. This information is consistent with content broadcasted in DTT.

TABLE 3. Test device specifications.

Resource	Specification
Device	MSI
Processor	12th Gen Intel(R) Core (TM) i7-12700H 2.70 GHz
Installed RAM	32.0 GB (31.7 GB usable)
System type	64-bit operating system, x64-based processor
Windows specifications	Windows 11 Pro

TABLE 4. Feature extraction time per video metric.

Video metric	Time (s)
Spatial Information	0.046365
Temporal Information	0.025019
Blurring	0.020148
Brightness	0.008143
Contrast	0.044653
Ringing	0.010312
Blockloss	0.156348
Blocking	0.232049
(All video metrics)	0.543037

C. TESTING DEVICE

The equipment used for the tests has the characteristics shown in Table 3.

With this device, using the tool described in subsection III-A with the set of more than 84000 individual images described in subsection III-B, the time it takes for the tool to perform the feature extraction is 0.543 s on average, per frame. The total time of the feature extraction in a 3-second measurement would be approximately 40.73 s, a value far from real-time processing.

Table 4 summarizes the time taken for the eight video metrics implemented for the feature extraction. Fig. 5 depicts the boxplot graphical representation with the same type of information. Blockloss and Blocking video metrics consume more than 71.5% of the time of all video metrics due to their computational cost.

D. TEST PLANNING

Measuring the computational cost of a computer process is not a simple task since many factors can change the performance of the device: running background processes, battery level, power savings options, memory level, temperature, etc. In the different tests and graphs, the computational cost information will not refer to the time but to the number of pixels processed in an image or the number of images processed in a 3-second measurement. Processing an image involves the feature extraction of that image.

In image resolution, a computational cost of 100% corresponds to processing the image at the original resolution of 1920×1080. In the number of images per measurement, a computational cost of 100% corresponds to processing the 75 images of the 3-second measurement. To process a 960×540 image would imply a computational cost of 25% (saving of 75%). To process 15 images per measurement would imply a computational cost of 20% (saving of 80%).

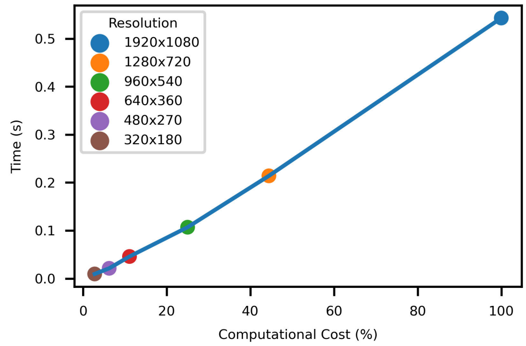

FIGURE 6. Graphical representation of the feature extraction time vs. video resolution.

The main objective of this work is to find the best approach that saves sufficient computational cost to allow the Video-MOS development tool to run in real-time on the test device, providing a MOS value estimation with the lowest possible error. For real-time execution on the test device, the computational cost must be below 7.37% and an acceptable MOS error value would be below 0.15, that is, below 3% due to a requirement set by content providers that use the Video-MOS quality probe. In addition, the findings will help to choose the approach that offers the best quality estimation accuracy to efficiently reduce the computational cost of the commercial solution if possible.

Section IV presents the results of applying different approaches exploring both spatial and temporal redundancy. For each approach, we provide the advantages, disadvantages, information about the computational cost, and the MOS error value obtained. Both values are obtained by comparing the processing in normal mode (complete image and all the images of the measurement) to each approach, using the 1123 test measures. The MOS error value is given in terms of mean absolute error (MAE).

Due to the time consumed in performing all the tests, section V presents an extensive and complete validation of the best approach using approximately 144 hours of audiovisual content from the six main HD DTT channels in Spain. Finally, section VI contains the main conclusions of this study.

IV. RESULTS

In this section we present the results obtained by applying different strategies based on the spatial redundancy and temporal redundancy of a video sequence. In the context of our study, results are presented following the methodology described in section III.

A. SPATIAL REDUNDANCY

The first set of approaches explores spatial redundancy by decreasing the image size. The study of analyzing how the feature extraction changes and how it affects the quality estimation is necessary using smaller image sizes. The proposed video resolutions maintain the 16:9 aspect ratio of the original size: 1280×720, 960×540, 480×270, 640×360 and 320×180.

TABLE 5. Time vs. Quality for the different interpolation methods of OpenCV resize function.

Interpolation Type	Time (s)	SSIM value
LINEAR	0.001519	0.921904
CUBIC	0.001617	0.925627
AREA	0.001882	0.900269
NEAREST	0.001372	0.900269
LANCZOS	0.002762	0.927619

Here we present two ideas: the first is to use a smaller area of the original image, and the second is to change the video resolution. For the second one, the OpenCV library provides the resize function and different methods to interpolate the pixel values: Linear, Cubic, Area, Nearest, and Lanczos [121]. The Cubic interpolation method has been the type chosen to make all the resolution changes. This choice is based on the results obtained after testing the different methods with all individual test images, seeking a compromise between the resizing time and the quality offered by each type of interpolation. The quality is measured by the SSIM FR IQA [41]. This metric has been widely used because of its simplicity and good results obtained in comparative studies between different metrics [122], [123]. SSIM is based on measuring the similarities of luminance, contrast, and structure between the reference and the distorted image. The metric is correlated with the visual perception of the HVS, and it is easily interpretable since the result of the comparison is normalized from 0 to 1. A SSIM value of 1 indicates a complete similarity between images, and lower values imply more distortion or difference between the images. Table 5 shows the results obtained for each type of interpolation by doing a double resizing process to 480×270 and to 1920×1080. SSIM calculation is performed at 1920×1080 resolution, comparing the original image with the one obtained after the two resizing processes.

The time taken to change the resolution to 480×270 for the Cubic method, on average between all the images, is 1.617 ms per image. The value is negligible when compared to the 0.543 s it takes for feature extraction per frame.

A significant reduction of the computational cost is achieved by performing the feature extraction on smaller images. Fig. 6 shows the feature extraction time according to the image resolution. The trend of the graph depicts an almost linear relationship between time and video resolution.

1) SPECIFIC AREA OF THE ORIGINAL IMAGE

The choice of which part of the image to use for quality estimation is not a simple decision. One option would be to use saliency detection to select the area of interest that would attract the attention of the end-users. However, saliency detection involves an additional and expensive computational cost due to the use of models for object detection, bright and contrasting area identification, motion estimation, and optical flow. For this reason, this approach uses only the central area of the image at different sizes for all proposed video resolutions. In many cases, the center of the image will contain the area of interest. Fig. 7 illustrates the graph between the computational cost and the MOS error value obtained for this approach selecting the central area of the original image.

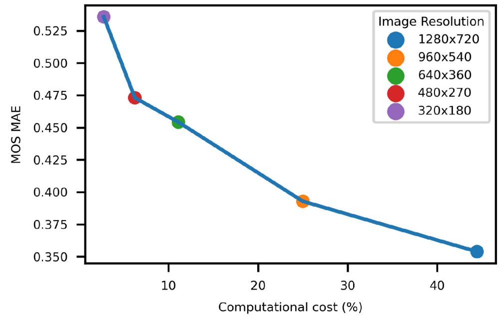

FIGURE 7. Graphical representation of the computational cost vs. MOS error in the central area of the original image approach.

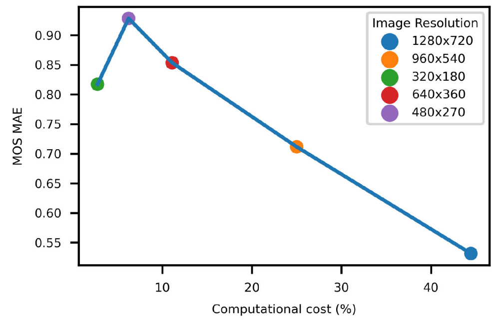

FIGURE 8. Graphical representation of the computational cost vs. MOS error in change of resolution approach.

The findings show a clear conclusion: the larger the central area, the lower the MOS error value. Selecting a specific area implies not processing part of the image and therefore not using that information in the quality estimation. If the characteristics of the unprocessed portion of the image are different from the characteristics of the central area, the feature vector will change and affect the MOS estimation. In terms of MOS error value, the approach does not offer good results since the error is 0.3538 at 1280×720 resolution.

2) CHANGE OF RESOLUTION

A change of resolution implies a subsampling of the image pixels. Although the information on the original image is maintained in terms of pixel values, subsampling involves a loss of high frequencies, blurring, a lower level of detail, and a change in image structure and edge information. The findings in terms of MOS error value are even worse than the previous approach. Fig. 8 shows the graph between the computational cost and the MOS error value for this approach making a resolution change with the Cubic method. The error is 0.5316 at 1280×720 resolution.

The analysis of the data shows considerable differences in feature extraction information at different image sizes. Video metrics that make use of edge information, high frequencies, and 3×3 fixed-size filters, such as Sobel or Laplacian operators, offer different features when the video resolution changes. However, this fact does not occur in video metrics

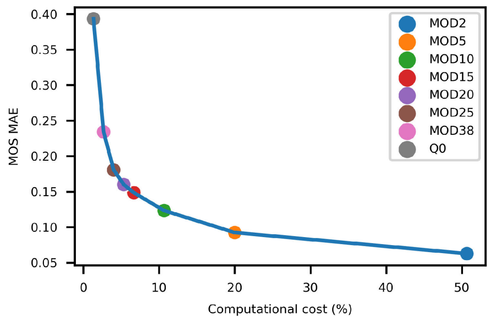

FIGURE 9. **Graphical representation of the computational cost vs. MOS error in uniform temporal sampling approach.**

that use only pixel-value information, since the subsampling process takes into account the value of all pixels of the original image.

In our hybrid NR-VQA solution, there are three pixel-value video metrics: Brightness, Contrast, and Temporal Information. If the change of resolution is applied only to the input images of these three metrics, keeping the original video resolution for the rest of the metrics, the MOS error value obtained is 0.0377 for 320 × 180 low resolution. For 960 × 540 and 480 × 270, the MOS errors are 0.0318 and 0.0395 respectively.

The time it takes now for feature extraction per frame using original and low image resolution goes from 0.543 to 0.466 s. The reduction of 77 ms per frame and a MOS error value below 0.04 make it a valid approach.

B. TEMPORAL REDUNDANCY

The easiest way to exploit the temporal redundancy is to apply a uniform temporal sampling and process only specific frames. The proposed temporal sampling modes are: MOD2, MOD5, MOD10, MOD15, MOD20, MOD25, MOD38, and QO. In MODX, X represents the distance between two consecutive processed images. Therefore, MOD15 indicates that one image is processed every fifteen frames. Thus, in a 3-second measurement, only five images would be processed with a computational cost for this mode of 6.68% of the original cost. Q0 indicates that only the first frame of the measurement is processed.

To maintain the correct performance of the solution, unprocessed frames keep the same features as the last processed one. This decision assumes that an unprocessed frame is identical to the last processed. Another decision taken is to always process the first frame of the measurement to guarantee that at least one frame is processed in the 3-second interval, regardless of the original frame rate.

With the idea of being able to use a longer uniform temporal sampling, we propose two additional mechanisms to force the processing of specific frames, by using the SSIM FR metric and the frame type at the video encoding level.

1) UNIFORM TEMPORAL SAMPLING

In uniform temporal sampling, a fixed number of images will always be processed depending on the selected mode.

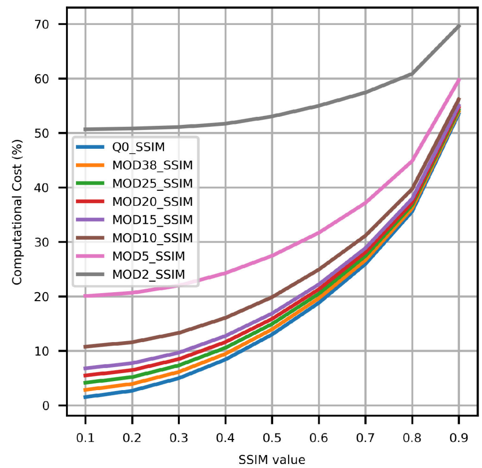

FIGURE 10. **Graphical representation of the SSIM threshold value vs. computational cost per each mode in uniform temporal sampling and SSIM mechanism approach.**

MOD15 always involves processing five images (assuming the same frame rate at 25 fps) regardless of the characteristics of the measurement and the variability between frames. For some measurements, these five images may be enough, for others, it may be either too many or too few depending on the complexity of the measurement. However, the main advantage of using a mode with a fixed number of images is that, by selecting a mode that works in real-time, the solution will always work in real-time since the computational cost will never be exceeded.

Figure 9 represents the graph between the computational cost and the MOS error value for the uniform temporal sampling approach. The curve depicts a decreasing logarithmic trend, where the MOS error decreases as the number of processed images increases.

For this case, MOD15 would be the mode chosen in this uniform temporal sampling approach since it is the mode with the lowest MOS error value that would allow achieving real-time. This mode would always process five images per measurement. It implies a computational cost of 6.68% (saving of 93.32%) with a MOS error of 0.1484.

2) UNIFORM TEMPORAL SAMPLING AND SSIM MECHANISM

This approach introduces the use of the SSIM mechanism in the uniform temporal sampling solution. This metric compares each image within the measurement with the previous processed one, activating the feature extraction in the frame if the difference is considerable. The idea with this mechanism is to use longer temporal sampling that sets fewer fixed frames and uses the SSIM metric to detect significant changes between frames. The first frame of the measurement is always processed, and the Temporal Information video metric must be computed within the next frame to a processed frame by the SSIM condition, since TI has to be computed between two adjacent frames: if Temporal Information is not computed, the change between frames would be maintained in the consecutive frames.

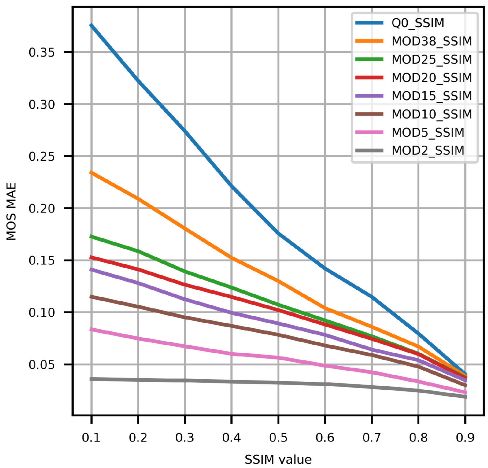

FIGURE 11. Graphical representation of the SSIM threshold value vs. MOS error per each mode in uniform temporal sampling and SSIM mechanism approach.

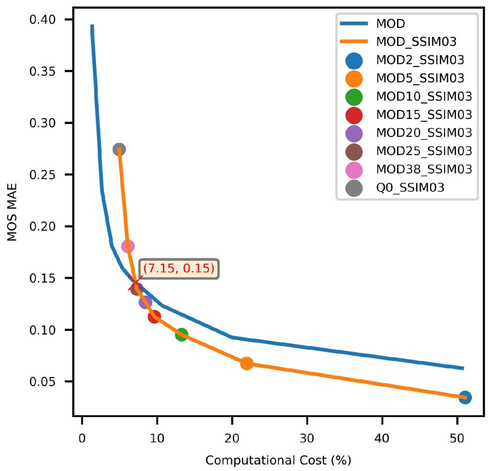

FIGURE 12. Graphical representation of the computational cost vs. MOS error in uniform temporal sampling and SSIM mechanism approach.

This approach has two drawbacks: the additional cost of computing the SSIM on all images of the measurement and the selection of a fixed SSIM threshold that determines the level of similarity needed to discard computation of the frames.

The test performed with all individual test images shows a high SSIM cost that increases with the image size. For the different resolutions, the SSIM temporal cost on average per image is 290 ms at 1920 × 1080, 135 ms at 1280 × 720, 73 ms at 960 × 540, 32 ms at 640 × 360, 14 ms at 480 × 270 and 4.6 ms at 320 × 180. For the lowest video resolution, SSIM cost is 4.6 ms, being 345 ms for the whole 3-second measurement.

The SSIM threshold for change detection will determine the number of images to be processed and thus affect the computational cost of the approach. A low threshold will allow the processing of a smaller number of images but will only detect significant changes between images. On the other hand, a high SSIM value would imply an excessive computational cost in the approach. Fig. 10 and Fig. 11 show the results obtained from applying nine different SSIM values from 0.1 to 0.9 for the eight uniform temporal sampling modes, in terms of computational cost and MOS error values. The SSIM threshold selected for the approach is 0.3 by seeking a compromise between the computational cost and the MOS error. In general terms, an SSIM value of 0.3 achieves real-time performance and a MOS error below 0.15.

Figure 12 represents the graph between the computational cost and the MOS error in the uniform temporal sampling approach with the SSIM threshold at 0.3. The figure also includes the uniform temporal sampling curve to establish a reference. The approach with the SSIM mechanism offers better results when the computational cost is greater than 7.15%.

The selected mode for this approach improving to uniform temporal sampling solution is MOD25_SSIM03 with a computational cost of 7.33% (computational saving of 92.67%) and MOS error of 0.1392. However, although on average the mode would allow real-time operation, the large variation in the number of images processed per measurement means that the mode is not valid in all situations, depending on the complexity and variability of the sequence. The number of images processed per measurement in this mode is 5.487 images on average, with a standard deviation of 6.1359.

MOD25_SSIM03 would not work in real-time in 17.36% of the test measurements because it would exceed the computational cost. To ensure real-time in all measurements, we propose MOD25_SSIM03_LIM, a limited version of MOD25_SSIM03 which stops processing frames when the maximum computational cost for real-time is reached, for each 3-second measure. MOD25_SSIM03_LIM implies a computational cost of 4.94% (computational saving of 95.06%) with a MOS error value of 0.1577.

For the 17.36% of that set of measurements, where the computational cost between MOD25_SSIM03_LIM and MOD15 is the same, the MOS errors obtained are 0.2194 and 0.1759 respectively.

In spite of the efforts, the several disadvantages of using the SSIM mechanism and the additional cost of metric calculation it carries, combined with the better results obtained with MOD15 for a significant percentage of measurements, make MOD25_SSIM03_LIM not a feasible approach.

3) UNIFORM TEMPORAL SAMPLING AND FRAME TYPE MECHANISM

This approach changes the SSIM mechanism for the frame type at the video encoding level. H.264/AVC video encoders (used in Spain in DTT HD broadcasted signal) use three types of frames for the video coding: I (Intra), P (Predictive), and B (Bi-directional). I frames are coded using only intra-frame prediction and are used as references for P and B frames prediction. P and B frames are coded using inter-frame prediction. However, P frames use only past frames as reference. B frames use both past and future frames.

H.264/AVC video encoders can use a static size or an adaptive structure for the GOP (Group of Pictures) to encode

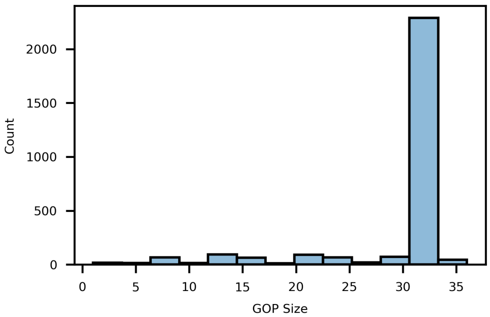

FIGURE 13. Test sequences. GOP size distribution.

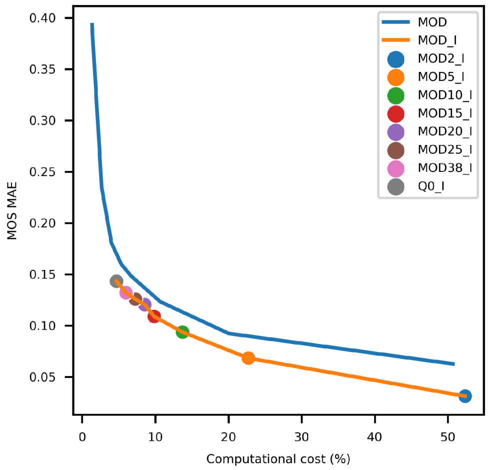

FIGURE 14. Graphical representation of the computational cost vs. MOS error in uniform temporal sampling and frame type mechanism approach.

the video. Adaptive GOP structure reacts better to scene changes and large variations in consecutive frames when generating predictions. In cases where a scene change is detected, in adaptive GOPs structures, video encoders can introduce an I frame [124], [125]. The assignment of the frame type and the GOP size plays a very important role in the encoding performance in terms of compression and quality.

Since I frames are often introduced in scene changes, these frames can be associated with low temporal redundancy instants. Similarly, P or B frames are intrinsically related to low temporal information. Therefore, arguably, in a generic situation, similar information can be obtained just by looking into the GOP structure rather than computing the SSIM algorithm. The idea of this approach is to focus the computation effort only on I frames, assuming they will have a lower SSIM value than P or B frames.

The reading of the metadata for obtaining the frame type is instantaneous and does not involve any additional computational cost. However, the main problem with the approach would be the appearance of small GOPs in video encoding. Too many I frames in a 3-second measurement could exceed the maximum computational cost and the approach would not work in real-time.

The frame type analysis in the 1123 test measurements reveals that there is an average of 2.56 I frames, 15.96 P frames, and 56.36 B frames per measurement. The average

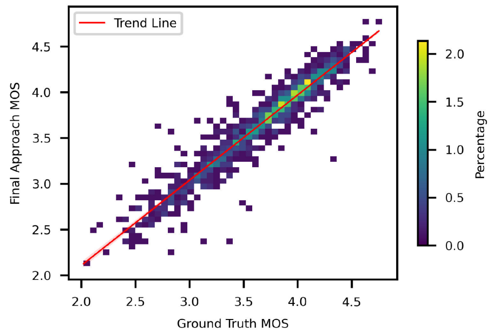

FIGURE 15. 2D Histogram of MOS values for test sequences: ground truth vs. proposed final approach. Trend line fitted with Linear Regression. Bin density is encoded using color.

of 2.56 I frames makes it possible to always try to process all I frames over the 3-second interval. Fig. 13 represents the GOP size distribution of the test set. For a total of 2865 GOPs, 66.67% have the IBBBP structure, M=4 and N=32. M indicates the distance between I and P frames or the distance between two consecutive P frames. N indicates the GOP size or the distance between two I frames.

The approach with frame type mechanism always processes the first frame of the measurement. Similar to the SSIM approach, the Temporal Information video metric is also computed in the frame following an I frame, since this frame type may indicate a change of scene. Fig. 14 shows the graph between the computational cost and the MOS error value in the uniform temporal sampling approach including the feature extraction also in the I frames. The graph includes the uniform temporal sampling curve to establish the reference.

The use of I frames in feature extraction improves the results offered by the uniform temporal sampling approach regardless of the computational cost. For the same number of processed images, the use of I frames offers a lower MOS error value. It is also possible to choose the longest temporal sampling of the proposed ones. Q0_I provides excellent results with a computational cost of 4.70% and a MOS error value of 0.1431.

Q0_I would not work in real-time for only the 0.53% of the test measurements, a percentage much lower than the obtained with the SSIM approach. To guarantee real-time also in that set of measurements, we propose the mode Q0_I_LIM, a limited version of Q0_I. Q0_I_LIM implies a computational cost of 4.68% with a MOS error of 0.1436.

Although for 100% of the measurements, the Q0_I_LIM performance is much better than MOD15, it is true that for that small set of 0.53% of the measurements, the results offered by MOD15 are better than Q0_I_LIM (MOS error of 0.2283 vs. 0.3617 respectively).

C. SPATIAL AND TEMPORAL REDUNDANCY

Finally, we summarize the lessons learned from exploring the spatial and temporal redundancies, and we combine

TABLE 6. Computational cost and MOS error for each approach. 100% of test measurements.

Approach	Computational cost AVG (%)	Computational cost STD (%)	MOS MAE
MOD15_SR	6.67	0.11	0.0864
Q0_I_LIM_SR	4.68	0.74	0.0924

TABLE 7. Computational cost and MOS error for each approach. 0.53% of test measurements.

Approach	Computational cost AVG (%)	Computational cost STD (%)	MOS MAE
MOD15_SR	6.67	0	0.17
Q0_I_LIM_SR	6.67	0	0.1783

the approaches that provided the best results in the tests performed.

Exploring the spatial redundancy of a video sequence, we propose to process the pixel-value video metrics at 320 × 180 low resolution, keeping the original resolution for the rest of the metrics. The processing time for Brightness, Contrast, and Temporal Information video metrics at low resolution is 1.21 ms per frame, that is 90.82 ms for a 3-second measurement. On the other hand, there is a saving of 77 ms per frame when processing these three metrics at low resolution. Data show that it is worth processing the three metrics in all frames at low resolution.

Exploring the temporal redundancy of a video sequence, we see the need to maintain MOD15 and Q0_I_LIM modes to guarantee real-time performance in all measurements. Although general findings show better performance of Q0_I_LIM, for complex sequences MOD15 offers better results.

Exploring both spatial and temporal redundancy of a video sequence, we propose the modes MOD15_SR and Q0_I_LIM_SR which are a combination of the techniques described above (SR in the name of the modes indicates Spatial Redundancy). Table 6 and Table 7 illustrate the results obtained for the complete set of the test sequences and for the set of complex sequences representing 0.53% of all, respectively. For the set of 0.53% of the measurements, both approaches offer the same results in terms of computational cost and MOS error. However, for the 100% of the measurements, for similar MOS errors below 0.1, Q0_I_LIM_SR implies much less computational cost than MOD15_SR.

Q0_I_LIM_SR is our final proposal, an approach that guarantees real-time in all measurements and achieves with the test sequences a computational cost of 4.68% and a MOS error value of 0.0924. Therefore, the computational cost saving is 95.32% with a MOS error below 0.1. Fig. 15 shows the ground truth of our proposal with the results of the solution in normal operation.

V. DISCUSSION AND VALIDATION

This section contains an exhaustive validation for the final selected proposal: Q0_I_LIM_SR. To guarantee the correct operation of this real-time approach in any possible scenario,

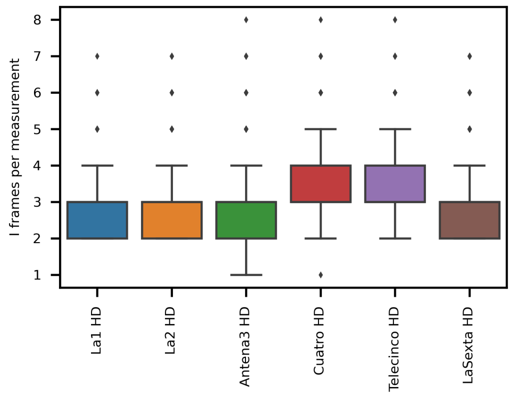

FIGURE 16. Validation sequences. I frames per measurement distribution.

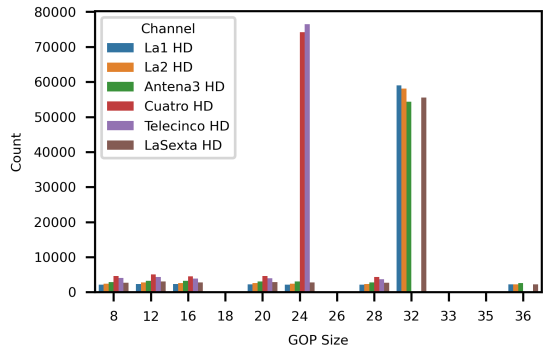

FIGURE 17. Validation sequences. GOP size distribution.

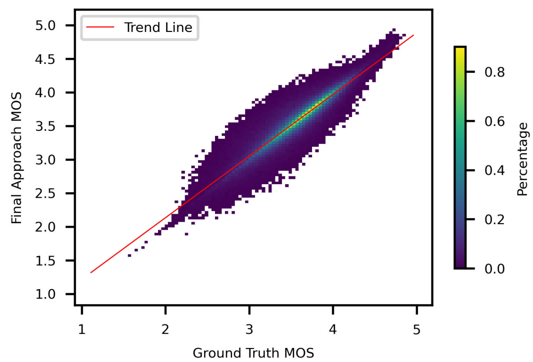

FIGURE 18. 2D Histogram of MOS values for validation sequences: ground truth vs. proposed final approach. Trend line fitted with Linear Regression. Bin density is encoded using color.

we have used six public Spanish DTT contents of 24 hours of duration from six of the most important HD channels in Spain: La1 HD, La2 HD, Antena3 HD, Cuatro HD, Telecinco HD and LaSexta HD. The 144 hours of audiovisual content and the diversity, both in terms of type of content (news, sports, musicals, documentaries, movies, series, etc.) and broadcasters, ensure an extensive validation of the final approach.

Figure 16 and Fig. 17 summarize the data analysis in terms of the number of I frames per measurement and the GOP size distribution for each content. The findings are similar to those obtained with the test sequences, which guarantees the validity of the use of I frames in the final approach for real-time operation. The most repeated GOP sizes are 32 images on channels La1 HD, La2 HD, Antena3 HD and LaSexta HD; and 24 images on channels Cuatro HD and Telecinco HD. Furthermore, analysis of the data shows a clear predominance of these GOP sizes depending on the channel. In terms of percentage with respect to the total number of GOPs per content, the GOP size of 32 images is repeated in 79% of the GOPs of La1 HD, 77% of La2 HD, 72% of Antena3 HD and 74% of LaSexta HD. In the same way, the GOP size of 24 images is repeated in 76% of the GOPs of Cuatro HD and in 79% of Telecinco HD. These GOP sizes of 32 and 24 images guarantee an average of 2.34 and 3.13 I frames, respectively, in a 3-second measurement of a DTT content.

In terms of computational cost and MOS error value, grouping the contents of the six HD channels, for a total of 174085 measurements, Q0_I_LIM_SR involves a computational cost of 5.04% (saving of 94.96%), with a standard deviation of the computational cost of 0.86%, and a MOS mean absolute error value of 0.1144. Fig. 18 represents the ground truth of the approach with all the validation measurements.

The promising results obtained in this validation with more than 144 hours of varied DTT content demonstrate the validity of the proposed solution with significant savings in computational cost and accuracy in quality estimation, for the NR-VQA model tested in our study, using both image downsampling technique for some video metrics and uniform temporal sampling technique with I frames. Due to the typical GOP size characteristics of HD DTT channels in Spain, our strategies are appropriate regardless the type of content, channel and broadcaster.

VI. CONCLUSION

With the big social impact of DTT TV in some countries, such as Spain, and the trend of increasing video IP traffic due to the multitude of audiovisual content, streaming services, social networks, and new consumption habits, the automatic estimation of perceived quality has become an interesting field of research. VQA has been studied for many years and a wide variety of different techniques exist in the literature. Discarded the subjective assessment for not being valid for real-time applications because of their complex methodologies and experiments with real observers, the NR objective metrics would be the most promising alternative for perceived quality estimation in real-time video streaming applications in the absence of the reference in most of the cases.

A complete revision of different models has been done in this paper, from traditional techniques to the most recent learning-based and deep-learning approaches. There are many challenges in NR-VQA with the emergence of new types of audiovisual content and the need to optimize the computational cost of the models due to the promising new audiovisual formats which involve much more information.

Motivated by all this, we have presented in this paper a proposal for computational cost reduction in video processing for QoE estimation, making use of the Video-MOS quality probe. The proposal can also be applied to other IQA/VQA measurement proposals. After exploring spatial and temporal redundancy with the objective of processing smaller images and/or a smaller number of images per measurement, the proposed final approach combines the video metrics feature extraction at both high and low video resolution along with a specific selection of frames based on a uniform temporal sampling and I frames. The test results for the final approach using 1123 measurements of HD content of DTT in Spain indicate a computational cost of 4.68% (computational cost saving of 95.32%) and MOS error value of 0.0924. The solution guarantees real-time operation on the test machine regardless of the complexity of the measurement.

The exhaustive validation of the proposed approach with more than 144 hours of video from six of the most important HD channels of DTT in Spain ensures the validity of the solution with the use of I frames, thanks to the typical GOP sizes used in H.264/AVC video encoding for HD content on DTT. For the more than 174000 3-second measurements used for the validation, the proposed approach involves a computational cost of 5.04% (cost saving of 94.96%) and a MOS mean absolute error value of 0.1144.

We believe that very promising findings have been obtained in this study, with significant savings in computational cost while maintaining high accuracy in MOS value estimation. Future research will address the use of new audiovisual formats, such as 4K and 8K video resolution and HFR (High Frame Rate) technology involving a higher number of images per second, that allows real-time operation in the commercial Video-MOS SaaS solution.

ACKNOWLEDGMENT

The authors would like to thank European company Video-MOS and the Chair of Video-MOS in UPM. They would also like to thank RTVE and the Chair of RTVE in UPM, for giving permission to use some screenshots of the test sequences in Fig. 2.

REFERENCES

[1] *Cisco Visual Networking Index: Forecast and Trends*, Cisco, San Jose, CA, USA, Nov. 17, 2023.

[2] *Growing App Complexity: Paving the Way for Digital Lifestyles and Immersive Experiences. Sandvine Phenomena*, Global Internet Phenomena Report, Waterloo, ON, Canada, Nov. 2023.

[3] N. Somraj, M. S. Kashi, S. P. Arun, and R. Soundararajan, "Understanding the perceived quality of video predictions," *Signal Process., Image Commun.*, vol. 102, Mar. 2022, Art. no. 116626, doi: 10.1016/j.image.2021.116626.

[4] *Vocabulary for Performance, Quality of Service and Quality of Experience*, document ITU-T P.10/G.100, 2017. [Online]. Available: https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-P.10-201711-I!!PDF-E&type=items

[5] Z. Akhtar, K. Siddique, A. Rattani, S. L. Lutfi, and T. H. Falk, "Why is multimedia quality of experience assessment a challenging problem?" *IEEE Access*, vol. 7, pp. 117897–117915, 2019, doi: 10.1109/ACCESS.2019.2936470.

[6] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document ITU-T P.910, 2022. [Online]. Available: https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-P.910-202207-S!!PDF-E&type=items

[7] *Methodologies for the Subjective Assessment of the Quality of Television Images*, document ITU-R Rec. BT.500, Accessed: Nov. 17, 2023. [Online]. Available: https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-15-202305-I!!PDF-E.pdf

[8] J. L. Dingquan, "Recent advances and challenges in video quality assessment," *ZTE Commun.*, vol. 17, no. 1, pp. 3–11, 2019.

[9] Y. Li, S. Meng, X. Zhang, M. Wang, S. Wang, Y. Wang, and S. Ma, "User-generated video quality assessment: A subjective and objective study," *IEEE Trans. Multimedia*, vol. 25, pp. 154–166, 2023, doi: 10.1109/TMM.2021.3122347.

[10] Y. Sugito and M. Bertalmío, "Performance evaluation of objective quality metrics on HLG-based HDR image coding," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Anaheim, CA, USA, Nov. 2018, pp. 96–100, doi: 10.1109/GLOBALSIP.2018.8646673. Accessed: Nov. 17, 2023.

[11] I. P. Gunawan, O. Cloramidina, S. B. Syafa'ah, R. H. Febriani, G. P. Kuntarto, and B. I. Santoso, "A review on high dynamic range (HDR) image quality assessment," *Int. J. Smart Sens. Intell. Syst.*, vol. 14, no. 1, pp. 1–17, Jan. 2021, doi: 10.21307/ijssis-2021-010.

[12] M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan, "Assessing visual quality of omnidirectional videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3516–3530, Dec. 2019, doi: 10.1109/TCSVT.2018.2886277.

[13] M. Xu, C. Li, S. Zhang, and P. L. Callet, "State-of-the-art in 360° video/image processing: Perception, assessment and compression," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 5–26, Jan. 2020, doi: 10.1109/JSTSP.2020.2966864.

[14] Y. Jin, M. Chen, T. Goodall, A. Patney, and A. C. Bovik, "Subjective and objective quality assessment of 2D and 3D foveated video compression in virtual reality," *IEEE Trans. Image Process.*, vol. 30, pp. 5905–5919, 2021, doi: 10.1109/TIP.2021.3087322.

[15] M. S. Anwar, J. Wang, W. Khan, A. Ullah, S. Ahmad, and Z. Fei, "Subjective QoE of 360-degree virtual reality videos and machine learning predictions," *IEEE Access*, vol. 8, pp. 148084–148099, 2020, doi: 10.1109/ACCESS.2020.3015556.

[16] N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, and S. Möller, "An objective and subjective quality assessment study of passive gaming video streaming," *Int. J. Netw. Manage.*, vol. 30, no. 3, pp. 1–16, May 2020, doi: 10.1002/nem.2054.

[17] Y. Gao, Y. Cao, T. Kou, W. Sun, Y. Dong, X. Liu, X. Min, and G. Zhai, "VDPVE: VQA dataset for perceptual video enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Vancouver, BC, Canada, Jun. 2023, doi: 10.1109/cvprw59228.2023.00152.

[18] X. Jiang, H. Yao, S. Zhang, X. Lu, and W. Zeng, "Night video enhancement using improved dark channel prior," in *Proc. IEEE Int. Conf. Image Process.*, Melbourne, QC, Australia, Sep. 2013, pp. 553–557, doi: 10.1109/ICIP.2013.6738114.

[19] P. Da, G. Song, P. Shi, and H. Zhang, "Perceptual quality assessment of nighttime video," *Displays*, vol. 70, Dec. 2021, Art. no. 102092, doi: 10.1016/j.displa.2021.102092.

[20] M. Nilsson, "Ultra high definition video formats and standardisation," BT Media Broadcast Res. Paper, London, U.K., Tech. Rep. Version 1.0, 2015.

[21] M. Cheon and J.-S. Lee, "Subjective and objective quality assessment of compressed 4K UHD videos for immersive experience," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 7, pp. 1467–1480, Jul. 2018, doi: 10.1109/TCSVT.2017.2683504.

[22] C. Bonnineau, W. Hamidouche, J. Fournier, N. Sidaty, J.-F. Travers, and O. Déforges, "Perceptual quality assessment of HEVC and VVC standards for 8K video," *IEEE Trans. Broadcast.*, vol. 68, no. 1, pp. 246–253, Mar. 2022, doi: 10.1109/TBC.2022.3140710.

[23] *Parameter Values for Ultra-high Definition Television Systems for Production and International Programme Exchange*, document ITU-R BT.2020, 2023. [Online]. Available: https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.2020-2-201510-I!!PDF-E.pdf

[24] *Parameter Values for the HDTV Standards for Production and International Programme Exchange*, document ITU-R BT.709, 2023. [Online]. Available: https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.709-6-201506-I!!PDF-E.pdf

[25] *Video-MOS | We Are the Video Content Quality Monitoring Experts*. Accessed: Nov. 17, 2023. [Online]. Available: https://www.video-mos.com/

[26] (2022). *Empowering Content Creators With AI Tools and Game With AI Tools and Game Engines Engine, Document Tech-I, Media Technology and Innovation,*. Accessed: Nov. 17, 2023. [Online]. Available: https://tech.ebu.ch/files/live/sites/tech/files/shared/tech-i/tech-i-051.pdf

[27] TM Broadcast. (2021). *RTVE Experimenta Con Vídeo-MOS La Automatización De La Evaluación De La Experiencia Del Espectador*. Accessed: Nov. 17, 2023. [Online]. Available: https://tmbroadcast.es/index.php/rtve-video-mos-analisis-resultados/

[28] D. Y. Lee, S. Paul, C. G. Bampis, H. Ko, J. Kim, S. Y. Jeong, B. Homan, and A. C. Bovik, "A subjective and objective study of space-time subsampled video quality," *IEEE Trans. Image Process.*, vol. 31, pp. 934–948, 2022, doi: 10.1109/TIP.2021.3137658. https://doi.org/10.1109/tip.2021.3137658

[29] J. Y. Lin, R. Song, C.-H. Wu, T. Liu, H. Wang, and C.-C.-J. Kuo, "MCL-V: A streaming video quality assessment database," *J. Vis. Commun. Image Represent.*, vol. 30, pp. 1–9, Jul. 2015, doi: 10.1016/j.jvcir.2015.02.012.

[30] A. Mackin, M. Afonso, F. Zhang, and D. Bull, "A study of subjective video quality at various spatial resolutions," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, Accessed: Nov. 17, 2023, pp. 2830–2834, doi: 10.1109/ICIP.2018.8451225.

[31] Q. Huang, S. Y. Jeong, S. Yang, D. Zhang, S. Hu, H. Y. Kim, J. S. Choi, and C.-C. J. Kuo, "Perceptual quality driven frame-rate selection (PQD-FRS) for high-frame-rate video," *IEEE Trans. Broadcast.*, vol. 62, no. 3, pp. 640–653, Sep. 2016, doi: 10.1109/TBC.2016.2570022.

[32] A. V. Katsenou, D. Ma, and D. R. Bull, "Perceptually-aligned frame rate selection using spatio-temporal features," in *Proc. Picture Coding Symp. (PCS)*, Jun. 2018, pp. 288–292, doi: 10.1109/PCS.2018.8456274.

[33] R. R. Ramachandra Rao, S. Göring, W. Robitza, B. Feiten, and A. Raake, "AVT-VQDB-UHD-1: A large scale video quality database for UHD-1," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, San Diego, CA, USA, Dec. 2019, pp. 17–177, doi: 10.1109/ISM46123.2019.00012.

[34] Y. Wang, Z. Chen, H. Jiang, S. Song, Y. Han, and G. Huang, "Adaptive focus for efficient video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 16229–16238, doi: 10.1109/ICCV48922.2021.01594.

[35] Z. Gao, G. Lu, and P. Yan, "Key-frame selection for video summarization: An approach of multidimensional time series analysis," *Multidimensional Syst. Signal Process.*, vol. 29, no. 4, pp. 1485–1505, Oct. 2018, doi: 10.1007/s11045-017-0513-9.

[36] J. Karotte and E. Sarma, "An evaluation of the effect of image downsampling on performance indicators of IQA algorithms," *ARPN J. Eng. Appl. Sci.*, vol. 10, pp. 7507–7513, Apr. 2015.

[37] C. Feng, D. Danier, F. Zhang, and D. Bull, "RankDVQA: Deep VQA based on ranking-inspired hybrid training," 2022, *arXiv:2202.08595*.

[38] M. Vranješ, S. Rimac-Drlje, and K. Grgić, "Review of objective video quality metrics and performance comparison using different databases," *Signal Process., Image Commun.*, vol. 28, no. 1, pp. 1–19, Jan. 2013, doi: 10.1016/j.image.2012.10.003.

[39] M. Vranjes, S. Rimac-Drlje, and D. Zagar, "Objective video quality metrics," in *Proc. ELMAR*, Zadar, Croatia, 2007, pp. 1–38, doi: 10.1109/elmar.2007.4418797.

[40] N. Ponomarenko, "On between-coefficient contrast masking of DCT basis functions," in *Proc. 3rd Int. Workshop Video Process. Quality Metrics Consum. Electron.*, 2007, pp. 1–4.

[41] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[42] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. Thrity-Seventh Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, 2003, pp. 1398–1402, doi: 10.1109/acssc.2003.1292216.

[43] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007, doi: 10.1109/tip.2007.901820.

[44] D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, Jan. 2010, Art. no. 011006, doi: 10.1117/1.3267105.

[45] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *Proc. 1st Int. Workshop Video Process. Quality Metrics Consum. Electron.*, 2005, vol. 7, no. 2, pp. 2117–2128.

[46] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011, doi: 10.1109/TIP.2011.2109730.

[47] S. Rimac-Drlje, M. Vranješ, and D. Žagar, "Foveated mean squared error—A novel video quality metric," *Multimedia Tools Appl.*, vol. 49, no. 3, pp. 425–445, Sep. 2010, doi: 10.1007/s11042-009-0442-1.

[48] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010, doi: 10.1109/TIP.2009.2034992.

[49] P. V. Vu and D. M. Chandler, "ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *J. Electron. Imag.*, vol. 23, no. 1, Feb. 2014, Art. no. 013016, doi: 10.1117/1.jei.23.1.013016.

[50] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *Proc. 18th IEEE Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 2505–2508, doi: 10.1109/ICIP.2011.6116171.

[51] F. Zhang and D. R. Bull, "A perception-based hybrid model for video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 6, pp. 1017–1028, Jun. 2016, doi: 10.1109/TCSVT.2015.2428551.

[52] K. Manasa and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2480–2492, Jun. 2016, doi: 10.1109/TIP.2016.2548247.

[53] J. Wu, Y. Liu, W. Dong, G. Shi, and W. Lin, "Quality assessment for video with degradation along salient trajectories," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2738–2749, Nov. 2019, doi: 10.1109/TMM.2019.2908377.

[54] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2256–2270, Aug. 2019, doi: 10.1109/TCSVT.2018.2868262.

[55] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "ST-GREED: Space-time generalized entropic differences for frame rate dependent video quality prediction," *IEEE Trans. Image Process.*, vol. 30, pp. 7446–7457, 2021, doi: 10.1109/TIP.2021.3106801.

[56] P. G. Freitas, W. Y. L. Akamine, and M. C. Q. Farias, "Using multiple spatio-temporal features to estimate video quality," *Signal Process., Image Commun.*, vol. 64, pp. 1–10, May 2018, doi: 10.1016/j.image.2018.02.010.

[57] S. V. R. Dendi, G. Krishnappa, and S. S. Channappayya, "Full-reference video quality assessment using deep 3D convolutional neural networks," in *Proc. Nat. Conf. Commun. (NCC)*, Bangalore, India, Feb. 2019, Accessed: Nov. 17, 2023, pp. 1–5, doi: 10.1109/NCC.2019.8732265.

[58] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *Computer Vision—ECCV 2018*. Cham, Switzerland: Springer, 2018, pp. 224–241. Accessed: Nov. 17, 2023, doi: 10.1007/978-3-030-01246-5_14.

[59] M. Xu, J. Chen, H. Wang, S. Liu, G. Li, and Z. Bai, "C3DVQA: full-reference video quality assessment with 3D convolutional neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 4447–4451, doi: 10.1109/ICASSP40776.2020.9053031.

[60] J. Chen, H. Wang, M. Xu, G. Li, and S. Liu, "Deep neural networks for end-to-end spatiotemporal video quality prediction and aggregation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Shenzhen, China, Jul. 2021, pp. 1–6, doi: 10.1109/ICME51207.2021.9428209.

[61] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018, doi: 10.1109/TIP.2017.2760518.

[62] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2022, doi: 10.1109/TPAMI.2020.3045810.

[63] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1969–1977, doi: 10.1109/CVPR.2017.213.

[64] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Image quality assessment using contrastive learning," *IEEE Trans. Image Process.*, vol. 31, pp. 4149–4161, 2022, doi: 10.1109/TIP.2022.3181496.

[65] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2013, doi: 10.1109/TCSVT.2012.2214933.

[66] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "SpEED-QA: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1333–1337, Sep. 2017, doi: 10.1109/LSP.2017.2726542.

[67] Z. M. Parvez Sazzad, Y. Kawayoke, and Y. Horita, "No reference image quality assessment for JPEG2000 based on spatial features," *Signal Process., Image Commun.*, vol. 23, no. 4, pp. 257–268, Apr. 2008, doi: 10.1016/j.image.2008.03.005.

[68] Z. Wang, A. C. Bovik, and B. L. Evan, "Blind measurement of blocking artifacts in images," in *Proc. Int. Conf. Image Process.*, Vancouver, BC, Canada, 2023, pp. 981–984, doi: 10.1109/ICIP.2000.899622.

[69] A. C. Bovik and S. Liu, "DCT-domain blind measurement of blocking artifacts in DCT-coded images," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Salt Lake City, UT, USA, Nov. 2023, pp. 1725–1728, doi: 10.1109/ICASSP.2001.941272.

[70] X. Zhu and P. Milanfar, "A no-reference sharpness metric sensitive to blur and noise," in *Proc. Int. Workshop Quality Multimedia Exper.*, San Diego, CA, USA, Jul. 2009, pp. 64–69, doi: 10.1109/QOMEX.2009.5246976.

[71] C. Chen, M. Izadi, and A. Kokaram, "A perceptual quality metric for videos distorted by spatially correlated noise," in *Proc. 24th ACM Int. Conf. Multimedia*, Amsterdam, The Netherlands. New York, NY, USA: ACM, Oct. 2016, pp. 1277–1285, doi: 10.1145/2964284.2964302.

[72] A. Norkin and N. Birkbeck, "Film grain synthesis for AV1 video codec," in *Proc. Data Compress. Conf.*, Snowbird, UT, USA, Mar. 2018, Accessed: Nov. 17, 2023, pp. 3–12, doi: 10.1109/DCC.2018.00008.

[73] H. Liu, N. Klomp, and I. Heynderickx, "A no-reference metric for perceived ringing artifacts in images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 4, pp. 529–539, Apr. 2010, doi: 10.1109/TCSVT.2009.2035848.

[74] X. Feng and J. P. Allebach, "Measurement of ringing artifacts in JPEG images," in *Proc. Electron. Imag.*, J. P. Allebach and H. Chao, Eds. Bellingham, WA, USA: SPIE, 2006, pp. 74–83, Accessed: Nov. 17, 2023, doi: 10.1117/12.645089.

[75] Z. Tu, J. Lin, Y. Wang, B. Adsumilli, and A. C. Bovik, "BBAND index: A no-reference banding artifact predictor," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 2712–2716, doi: 10.1109/ICASSP40776.2020.9053634.

[76] D. L. Ruderman, "The statistics of natural images," *Netw., Comput. Neural Syst.*, vol. 5, no. 4, pp. 517–548, Jan. 1994, Accessed: Nov. 17, 2023, doi: 10.1088/0954-898x_5_4_006.

[77] M. A. Saad, A. C. Bovik, and C. Charrier, "A DCT statistics-based blind image quality index," *IEEE Signal Process. Lett.*, vol. 17, no. 6, pp. 583–586, Jun. 2010, doi: 10.1109/LSP.2010.2045550.

[78] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012, doi: 10.1109/TIP.2012.2191563.

[79] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013, doi: 10.1109/lsp.2012.2227726.

[80] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012, doi: 10.1109/TIP.2012.2214050.

[81] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010, doi: 10.1109/LSP.2010.2043888.

[82] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011, doi: 10.1109/TIP.2011.2147325.

[83] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014, doi: 10.1109/TIP.2014.2355716.

[84] Y. Zhang and D. M. Chandler, "No-reference image quality assessment based on log-derivative statistics of natural scenes," *J. Electron. Imag.*, vol. 22, no. 4, Dec. 2013, Art. no. 043025, doi: 10.1117/1.jei.22.4.043025.

[85] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-reference image quality assessment for high dynamic range images," in *Proc. 50th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2016, pp. 1847–1852, doi: 10.1109/ACSSC.2016.7869704.

[86] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, no. 1, p. 32, Jan. 2017, doi: 10.1167/17.1.32.

[87] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 1098–1105, doi: 10.1109/CVPR.2012.6247789.
[88] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Trans. Image Process.*, vol. 30, pp. 4449–4464, 2021, doi: 10.1109/TIP.2021.3072221.
[89] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, Mar. 2014, doi: 10.1109/TIP.2014.2299154. https://doi.org/10.1109/tip.2014.2299154
[90] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity Oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, Jan. 2016, doi: 10.1109/TIP.2015.2502725.
[91] X. Li, Q. Guo, and X. Lu, "Spatiotemporal statistics for video quality assessment," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3329–3342, Jul. 2016, doi: 10.1109/TIP.2016.2568752. https://doi.org/10.1109/tip.2016.2568752
[92] H. Men, H. Lin, and D. Saupe, "Spatiotemporal feature combination model for no-reference video quality assessment," in *Proc. 10th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Cagliari, May 2018, Accessed: Nov. 17, 2023, pp. 1–3, doi: 10.1109/QOMEX.2018.8463426.
[93] J. P. Ebenezer, Z. Shang, Y. Wu, H. Wei, and A. C. Bovik, "No-reference video quality assessment using space-time chips," in *Proc. IEEE 22nd Int. Workshop Multimedia Signal Process. (MMSP)*, Tampere, Sep. 2020, Accessed: Nov. 17, 2023, pp. 1–6, doi: 10.1109/MMSP48831.2020.9287151.
[94] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5923–5938, Dec. 2019, doi: 10.1109/TIP.2019.2923051.
[95] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-VQ: 'Patching up' the video quality problem," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 14014–14024, doi: 10.1109/CVPR46437.2021.01380.
[96] D. Varga, "No-reference video quality assessment using multi-pooled, saliency weighted deep features and decision fusion," *Sensors*, vol. 22, no. 6, p. 2209, Mar. 2022, doi: 10.3390/s22062209.
[97] B. Chen, L. Zhu, G. Li, F. Lu, H. Fan, and S. Wang, "Learning generalized spatial–temporal deep feature representation for no-reference video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1903–1916, Apr. 2022, doi: 10.1109/TCSVT.2021.3088505.
[98] W. Zhou and Z. Chen, "Deep local and global spatiotemporal feature aggregation for blind video quality assessment," in *Proc. IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, Macau, Dec. 2020, pp. 338–341, doi: 10.1109/VCIP49819.2020.9301764.
[99] J. Korhonen, Y. Su, and J. You, "Blind natural video quality prediction via statistical temporal features and deep spatial features," in *Proc. 28th ACM Int. Conf. Multimedia*. New York, NY, USA: ACM, Oct. 2020, pp. 3311–3319, doi: 10.1145/3394171.3413845.
[100] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proc. 27th ACM Int. Conf. Multimedia*, France. New York, NY, USA: ACM, Oct. 2019, pp. 2351–2359, doi: 10.1145/3343031.3351028.
[101] D. Li, T. Jiang, and M. Jiang, "Unified quality assessment of in-the-Wild videos with mixed datasets training," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1238–1257, Apr. 2021, doi: 10.1007/s11263-020-01408-w.
[102] H. Wu, C. Chen, L. Liao, J. Hou, W. Sun, Q. Yan, and W. Lin, "DisCoVQA: Temporal distortion-content transformers for video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4840–4854, Nov. 2023, doi: 10.1109/TCSVT.2023.3249741.
[103] Y. Wang, J. Ke, H. Talebi, J. G. Yim, N. Birkbeck, B. Adsumilli, P. Milanfar, and F. Yang, "Rich features for perceptual quality assessment of UGC videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 13430–13439, doi: 10.1109/CVPR46437.2021.01323.
[104] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang, "Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 5944–5958, Sep. 2022, doi: 10.1109/TCSVT.2022.3164467.
[105] W. Liu, Z. Duanmu, and Z. Wang, "End-to-end blind quality assessment of compressed videos using deep neural networks," in *Proc. 26th ACM Int. Conf. Multimedia*. New York, NY, USA: ACM, Oct. 2018, pp. 546–554, doi: 10.1145/3240508.3240643.
[106] L. Lin, Z. Wang, J. He, W. Chen, Y. Xu, and T. Zhao, "Deep quality assessment of compressed videos: A subjective and objective study," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2616–2626, Jun. 2022, doi: 10.1109/TCSVT.2022.3227039.
[107] Y. Li, L.-M. Po, C.-H. Cheung, X. Xu, L. Feng, F. Yuan, and K.-W. Cheung, "No-reference video quality assessment with 3D shearlet transform and convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 6, pp. 1044–1057, Jun. 2016, doi: 10.1109/TCSVT.2015.2430711.
[108] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "RAPIQUE: Rapid and accurate video quality prediction of user generated content," *IEEE Open J. Signal Process.*, vol. 2, pp. 425–440, Nov. 2021, doi: 10.1109/OJSP.2021.3090333.
[109] M. Agarla, L. Celona, and R. Schettini, "An efficient method for no-reference video quality assessment," *J. Imag.*, vol. 7, no. 3, p. 55, Mar. 2021, doi: 10.3390/jimaging7030055.
[110] Y. Fang, Z. Li, J. Yan, X. Sui, and H. Liu, "Study of spatio-temporal modeling in video quality assessment," *IEEE Trans. Image Process.*, vol. 32, pp. 2693–2702, 2023, doi: 10.1109/TIP.2023.3272480.
[111] A. K. Vishwakarma and K. M. Bhurchandi, "No-reference video quality assessment using local structural and quality-aware deep features," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023, doi: 10.1109/TIM.2023.3273654.
[112] J. Ke, T. Zhang, Y. Wang, P. Milanfar, and F. Yang, "MRET: multi-resolution transformer for video quality assessment," *Frontiers Signal Process.*, vol. 3, pp. 1–10, Mar. 2023, doi: 10.3389/frsip.2023.1137006.
[113] D. Varga, "No-reference video quality assessment using the temporal statistics of global and local image features," *Sensors*, vol. 22, no. 24, p. 9696, Dec. 2022, doi: 10.3390/s22249696.
[114] K. Zhao, K. Yuan, M. Sun, and X. Wen, "Zoom-VQA: Patches, frames and clips integration for video quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Vancouver, BC, Canada, Jun. 2023, pp. 1302–1310, doi: 10.1109/cvprw59228.2023.00137.
[115] H. Wu, "FAST-VQA: Efficient end-to-end video quality assessment with fragment sampling," in *Computer Vision—ECCV 2022* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2022, pp. 538–554.
[116] H. Wu, "Disentangling aesthetic and technical effects for video quality assessment of user generated content," 2022, *arXiv:/2211.04894*.
[117] *La UPM Y Video-Mos Se Unen Para Mejorar La Calidad De La Experiencia Del Consumidor De Contenidos Audiovisuales*. Accessed: Feb. 8, 2024. [Online]. Available: https://www.upm.es/upm?id=CON03229&prefmt=articulo&fmt=detail
[118] *Real Decreto 391/2019, De 21 De Junio, Por El Que Se Aprueba El Plan Técnico Nacional De La Televisión Digital Terrestre Y Se Regulan Determinados Aspectos Para La Liberación Del Segundo Dividendo Digital*, document BOE-A-2019-9513, Ministerio de Economía y Empresa, Gobierno de España, 2019. [Online]. Available: https://www.boe.es/eli/es/rd/2019/06/21/391
[119] *RTVE. Es*. Accessed: Nov. 17, 2023. [Online]. Available: https://www.rtve.es/
[120] *Cátedra RTVE En La UPM*. Accessed: Nov. 17, 2023. [Online]. Available: http://catedra.rtve.etsit.upm.es/
[121] *OpenCV*. Accessed: Nov. 17, 2023. [Online]. Available: https://opencv.org/
[122] J. Vlaovic, M. Vranješ, D. Grabic, and D. Samardžija, "Comparison of objective video quality assessment methods on videos with different spatial resolutions," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Osijek, Croatia, Jun. 2019, Accessed: Nov. 17, 2023, pp. 287–292, doi: 10.1109/IWSSIP.2019.8787324.
[123] U. Sara, M. Akter, and M. S. Uddin, "Image quality assessment through FSIM, SSIM, MSE and PSNR—A comparative study," *J. Comput. Commun.*, vol. 7, no. 3, pp. 8–18, 2019, doi: 10.4236/jcc.2019.73002.
[124] B. Zatt, M. Porto, J. Scharcanski, and S. Bampi, "Gop structure adaptive to the video content for efficient H.264/AVC encoding," in *Proc. IEEE Int. Conf. Image Process.*, Hong Kong, Sep. 2010, pp. 3053–3056, doi: 10.1109/ICIP.2010.5651700.
[125] L. Krulikovská, J. Polec, and M. Martinovic, "Adaptive group of pictures structure based on the positions of video cuts," *Int. J. Comput. Inf. Eng.*, vol. 7, no. 7, pp. 1–4, Jan. 2013.

ÁLVARO LLORENTE received the Bachelor of Engineering degree in telecommunication technologies and services and the master's degree in telecommunication engineering from Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree in communication technologies and systems with the Signals, Systems and Radiocommunications (SSR) Department. He has defended the final degree project entitled "Design and Execution of Subjective Quality Tests in Ultra High Definition TV." His master's degree project focused on the development of applications for a multi-device HbbTV 2.0.1 scenario with synchronous OTT services. Since 2015, he has been a Researcher with the SSR Department, collaborating in different national and international research projects and with the Chair of RTVE, UPM. His professional interests include the broadcast of digital television, accessibility services, interactive TV, new video and audio formats and technologies, and quality of experience in audiovisual signals.

JAVIER GUINEA PÉREZ received the B.S. degree in telecommunication engineering from Universidad Politécnica de Madrid, Madrid, Spain, in 2018, and the double M.Sc. degree in computer science from Kungliga Tekniska Hogskolan, Stockholm, Sweden, and the Delft University of Technology, Delft, The Netherlands, in 2020. Since 2021, he has been a Researcher with the Signals, Systems and Radiocommunications (SSR) Department, Universidad Politécnica de Madrid. Since 2023, he has been a Senior Data Scientist with Video-MOS, Madrid. His research interests include computer vision, deep learning, medical imaging, and natural language processing.

JUAN ANTONIO RODRIGO received the degree in telecommunication engineering and the master's and Ph.D. (cum laude) degrees in communications technologies and systems from Universidad Politécnica de Madrid, in 2007, 2010, and 2016, respectively. His thesis entitled "Obtaining and Adapting Depth Maps in Stereoscopic Video: Influence on Encoding and Three-Dimensional Visual Behavior in Real Time." He is a Lecturer of computer engineering with the Computer Systems, Universidad Politécnica de Madrid. Since February 2007, he has been a member of Grupo de Aplicación de Telecomunicaciones Visuales (GATV), Departmento de Señales, Sistemas y Radiodifusión, ETSIT. During these years, he has focused his professional work on the collaboration of R+D+i projects in different research areas that include real-time video processing, encoding for high-definition television (HDTV) and 3-D television (3DTV), video streaming, quality of experience, image and video analysis applied to artificial vision, machine learning, embedded programming for real-time video processing, and process automation. He has various publications in journals and international conferences endorse his research work.

DAVID JIMÉNEZ (Member, IEEE) received the degree (Hons.) in telecommunications engineer from ETS de Ingenieros de Telecomunicación, in 2004, and the Ph.D. degree (cum laude) in telecommunications engineering from Universidad Politécnica de Madrid. He is a Lecturer of electrical engineering with the Department of Physical Electronics, Electrical Engineering and Applied Physics, Universidad Politécnica de Madrid. Since 2004, he has been a member of Grupo de Aplicación de Telecomunicaciones Visuales (G@TV). His professional activity focuses on the field of smart energy grids (smart grids), the electric vehicles, 5G communications systems, and the application of IoT in the energy field. Lines of work that he shares with audiovisual technologies, video processing and treatment, video quality analysis, and the average quality of the user experience. He is a member of the Parity Commission of the Chair of RTVE, UPM. He participates in various large-scale national and European projects on topics, such as the evolution of 5G and beyond networks (NEMO and CODECO), the development of energy communities in Europe (eNEURON), and the analysis of hybrid power plants comprising hydropower, Li-ion batteries, and supercapacitors (HybridHydro).

JOSÉ MANUEL MENÉNDEZ (Senior Member, IEEE) was the Director of the Visual Telecommunication Application Research Group–GATV, from 2004 to 2019. He has been the Director of the Chair of the Spanish Public Broadcaster RTVE, Universidad Politécnica de Madrid (UPM), since 2015, where he is currently a Full Professor with the Signal, Systems and Radiocommunications Department, ETS Ingenieros de Telecomunicación, where he has been teaching on topics related to communications, audio-visual systems, and computer vision, since 1994. He has extensive experience in participating in and directing research projects (more than 200), both national and European, on topics related to visual communications, digital television, and computer vision. He has published more than 230 papers on these subjects in international journals and conferences and has been invited to give more than 80 invited lectures (both national and international). He is the author of three granted patents and 12 software registrations. He has been a Regular Reviewer of the IEEE Signal Processing Society, since 2000, for different international journals and conferences sponsored by that entity. He has been an Evaluator of the IET Society for several journals, since 2009. He also collaborates regularly with different national and regional entities and the European Commission in the certification of accredited laboratories, in the evaluation and review of R+D+i projects (FP-VI, FP-VII, H2020, and HE) and performs consulting work for broadcasters and companies in the telecommunications sector, Spain. He has participated in the creation of three technology-based spin-off companies and collaborated in the launch of three other companies in the information and communications technology sector.