

UNICA UNIVERSITÀ DEGLI STUDI DI CAGLIARI



This is the Author's *accepted* manuscript version of the following contribution:

Simone Porcu, Alessandro Floris, Jan-Niklas Voigt-Antons, Luigi Atzori, Sebastian Möller, "Estimation of the Quality of Experience During Video Streaming From Facial Expression and Gaze Direction" in *IEEE Transactions on Network and Service Management*, Volume 17, Issue 4, Article number 9171870, pp. 2702-2716.

© 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

The publisher's version is available at:

http://dx.doi.org/10.1109/TNSM.2020.3018303

When citing, please refer to the published version.

This full text was downloaded from UNICA IRIS https://iris.unica.it/

Estimation of the Quality of Experience during Video Streaming from Facial Expression and Gaze Direction

Simone Porcu, *Student Member, IEEE,* Alessandro Floris, *Member, IEEE,* Jan-Niklas Voigt-Antons, *Member, IEEE,* Luigi Atzori, *Senior Member, IEEE,* Sebastian Möller, *Senior Member, IEEE*

***Abstract*—This paper investigates the possibility to estimate the perceived Quality of Experience (QoE) automatically and unobtrusively by analyzing the face of the consumer of video streaming services, from which facial expression and gaze direction are extracted. If effective, this would be a valuable tool for the monitoring of personal QoE during video streaming services without asking the user to provide feedback, with great advantages for service management. Additionally, this would eliminate the bias of subjective tests and would avoid bothering the viewers with questions to collect opinions and feedback. The performed analysis relies on two different experiments: i) a crowdsourcing test, where the videos are subject to impairments caused by long initial delays and re-buffering events; ii) a laboratory test, where the videos are affected by blurring effects. The facial Action Units (AU) that represent the contractions of specific facial muscles together with the position of the eyes' pupils are extracted to identify the correlation between perceived quality and facial expressions. An SVM with a quadratic kernel and a k-NN classifier have been tested to predict the QoE from these features. These have also been combined with measured application-level parameters to improve the quality prediction. From the performed experiments, it results that the best performance is obtained with the k-NN classifier by combining all the described features and after training it with both the datasets, with a prediction accuracy as high as 93.9% outperforming the state of the art achievements.**

***Index Terms*—Video streaming, Quality of Experience, Facial expressions, Gaze direction, Machine learning, Video Key Quality Indicators, QoE estimation.**

I. Introduction

THE Internet traffic has grown considerably in recent years. The most relevant causes are the increased popularity of video streaming platforms, such as YouTube and Netflix, as well as the augmented sharing of multimedia contents (videos in particular) through social networks. Indeed, Cisco predicts global IP video traffic to be 82% of all IP traffic by 2022, up from 75% in 2017 [1]. Such an increase of multimedia traffic, together with the high service quality expected by end-users, shifted the research studies to user-centered network and service management approaches [2], [3]. Accordingly, the Quality of Experience (QoE) has become more and more important for the successful deployment of multimedia services as the QoE reflects the subjective quality perceived by the user [4]. QoE is defined as the 'the degree of delight or annoyance of the user of an application or service.' [5]. The perceived QoE can be usually assessed by objective or subjective measures. Objective methods are mathematical metrics classified into three main categories, namely, Full Reference (FR), Reduced Reference (RR), and No Reference (NR), based on the availability of the reference stimuli. These metrics compare the original and distorted data to define the quality. Subjective tests require humans to rate the perceived quality of presented stimuli, typically using discrete quality scales, from which the Mean Opinion Score (MOS) is derived [6].

The collection of user's subjective perceived quality and feedback is of paramount importance to identify the root causes of quality degradation and to take the necessary corrective actions in service and network management [7], [8]. Indeed, besides coding and network distortions that are measurable, there are other factors that may influence the user's experience, such as the user's characteristics, the device and the context of use, which may vary depending on the user [9], [10]. However, while convenient and effective, this kind of self-report technique can be problematic because it may be subject to bias from factors not related to the stimulus. These may include, for example, the interviewer's reaction to the questions, the way the questions are formulated, and the context (tests are typically conducted in the laboratory). Additionally, users may be annoyed by surveys and interviews, which are generally boring and time-consuming. For these reasons, alternative approaches for QoE evaluation started to be investigated, such as psychophysiological measures (e.g., electroencephalography (EEG), gaze direction) and facial expressions [11], [12].

The subject of this paper falls into this area, with a specific focus on the possibility to estimate the perceived QoE automatically and unobtrusively by analyzing the face of the video viewer, from which facial expression and gaze direction are extracted. If effective, this would be a valuable tool for the monitoring of personal QoE during video streaming services without asking the user to provide feedback, with great advantages for service management. Some studies have

This work was partially supported by Italian Ministry of University and Research (MIUR), within the Smart Cities framework (Project Netergit, ID: PON04a200490).

S. Porcu, A. Floris and L. Atzori are with the DIEE, University of Cagliari, UdR CNIT of Cagliari, Italy (e-mail: simone.porcu@unica.it; alessandro.floris84@unica.it; l.atzori@ieee.org).

J.-N. Voigt-Antons and S. Möller are with the Quality and Usability Lab, Technische Universität Berlin, Germany, and with the German Research Center for Artificial Intelligence (DFKI), Berlin, Germany (e-mail: jan-niklas.voigt-antons@tu-berlin.de; sebastian.moeller@tu-berlin.de).

started to investigate the correlation between facial expressions and the human's emotional state of users when consuming a multimedia service, such as watching a video sequence or during a VoIP call [12]–[14]. Therefore, we specifically focused on the estimation of perceived QoE for video streaming services. To this, we conducted two different experiments: i) a crowdsourcing test in which participants were asked to watch and rate the quality of 20 videos subject to impairments caused by combinations of 3 video Key Quality Indicators (KQIs), namely, initial delay, number of buffering events and duration of buffering events [15]; ii) a laboratory test in which participants were asked to watch and rate the quality of 105 videos subject to impairments caused by combinations of 2 video KQIs, namely, size of blurring kernel and blurring duration [16].

The facial Action Units (AUs) were considered as the features to capture the facial expression of the viewer, whereas the position of the eyes' pupils was the feature considered to evaluate the viewer's gaze direction. We define 3 facial metrics to derive significance from these facial features. Moreover, we define the impairment level (IL) feature, which has the objective to highlight the level of impairment introduced to the video by the combinations of video KQIs with respect to the video without impairments. The IL feature, together with the aforementioned facial metrics and the respective QoE values provided by the participants, were used to train different machine learning (ML) classifiers aimed at QoE estimation. Specifically, we considered two QoE estimation models: i) *AU&GDtoQoE*: takes as input the AU features and the position of the eyes' pupil; ii) *AU&GD&KQItoQoE*: takes as input the AU features, the position of the eyes' pupil and the service's KQIs. The performance of the considered ML classifiers was compared in terms of accuracy, sensitivity, and specificity, whereas the QoE estimation models were validated using three different quality scales: 5-level, 3-level, and 2-level scales.

The following are the main contributions of this article:

- We design a methodology to implement a QoE estimation system based on human facial expressions, gaze direction, and service's KQIs, which would be a key component in video streaming service management.
- We conceive a method to merge features generated from different experiments under different conditions, i.e., different video length, different KQIs.
- We compare the performance of the proposed QoE estimation models with state-of-the-art ML-based models based on facial expressions and network Quality of Service (QoS) features, outperforming their results.
- We compare the performance of the proposed QoE estimation models with state-of-the-art objective models for video quality, from which encouraging results are obtained.

The paper is structured as follows. Section II discusses the major related work in this area. Section III presents the proposed methodology while Section IV describes the conducted experiments for data collection. Section V presents the procedures followed to pre-process the collected data. In Section VI, the proposed ML-based QoE models are presented whereas Section VII shows the achieved results and compares them with state-of-the-art models. Finally, in Section VIII we draw the conclusions.

II. Related Work

Various QoE evaluation approaches, alternative to survey and interviews, have been proposed in the literature to support subjective tests and provide deeper insights into high-level QoE features. Most of these studies are based on psychophysiological measures, such as electroencephalography (EEG), to overcome the problem of potentially misleading rating scales and conscious decision making by identifying implicit responses to physical stimuli [11].

With regard to speech quality assessment using EEG, the studies in [17] and [18] investigated the potentials of event-related-potentials (ERP) analysis as a valid tool to indicate variation in quality perception. In particular, in [18], the research focused on the P300 component and its two subcomponents, P3a and P3b. With regard to audiovisual quality assessment using EEG, the study in [19] concluded that for longer sequences, low-quality conditions led to higher α and θ waves (the result of EEG analysis), which respectively indicate decreased alertness and attention. The α activity was also found to be significantly predictive of video quality [20]. In [21], α values and pupil dilation were used as parameters of a linear regression model, whose predictions of QoE scores achieved a correlation value of 0.64 with subjective QoE scores. Other studies considered eye-related measurements to investigate cognitive activities relevant to QoE assessment that are not easily observable through methods such as the EEG. For example, by monitoring the movement of the eyes, it is possible to collect valuable insights regarding visual attention. Eye blink rate is instead related to visual fatigue and pupil dilation to cognitive load. A gaze contingent display (GCD) was used by [22] to study the impact of spatio-temporal distortions in the peri foveal and extra-peri foveal regions using an eye-tracker. Eye-tracking data and the associated differential MOS (DMOS) results obtained from a subjective test involving videos containing localized distortions are examined in [23]. Moreover, studies on gaze tracking have shown that distortions located in salient regions have a significantly higher impact on quality perception as compared to distortions in non-salient regions [24]. For this reason, gaze direction is often integrated into image and video quality metrics with the objective to further improve their quality prediction performance [26].

There is a wide variety of related work when it comes to utilizing ML techniques to predict the QoE. ML has been utilized for video quality prediction in terms of perceptual quality [27], creation of predictive NR metrics for video quality assessment [28], encrypted video streaming [29], human activity recognition [30], video QoE in wireless cellular networks [31], [32] and QoE prediction for Voice over IP (VoIP) calls [33], [34]. Deep learning models are used in [35] for multimedia QoE prediction, in [36] for adaptive video streaming and in [37] for live video streaming. For passive gaming video streaming applications, NR video quality estimation based on ML was used by [38]. A review of predictive QoE management using ML in video streaming services is presented in [39].

TABLE I
COMPARISON OF REFERENCE STUDIES BASED ON PSYCHOPHYSIOLOGICAL MEASURES.

Reference	Stimuli	Distortion	Assessment	Psychophysiological measure
[12]	2D videos	Bandwidth, delay, video resolution	Subjective test	Face expression
[13]	2D videos	No	Subjective test	Face expression
[14]	Speech	Delay, bandwidth and packet loss	Subjective test	Acoustic feature
[15]	2D videos	Delay, buffering events	Subjective test	Face expression and eyetracking
[16]	2D videos	Blurring	Subjective test	Face expression and eyetracking
[17]	Speech	Signal-correlated noise	Subjective test	EEG
[18]	Speech	Signal-correlated noise	Subjective test	EEG
[19]	Audiovisual	Coding at low bitrate	Subjective test	EEG and EOG (electrooculogram)
[20]	2D and 3D videos	Coding at high quantization parameter	Subjective test	EEG
[21]	2D videos	Coding at low bitrate	Subjective test	EEG and eyetracking
[22]	2D videos	Coding at high quantization parameter	GCD and gaze analysis-techniques	Eyetracking
[23]	2D videos	Packet loss	Subjective test	Eyetracking
[24]	2D videos	Packet loss	Subjective test	Eyetracking
[25]	2D videos	Induced amusement and sadness emotions	Subjective test	Face expression and 15 physiological measures

The relevant techniques of sentiment analysis and Facial Expression Recognition (FER) have also been recently explored by following ML-based approaches. In [40], a CNN is combined with specific image pre-processing steps aimed to extract only expression specific features from a face image and explore the presentation order of the samples during training. [41] proposes a novel deep-based framework, which mainly consists of two branches of the CNN. One branch extracts local features from image patches while the other extracts holistic features from the whole expressional image. Local and holistic features are then aggregated before the classification process. [42] built an emotion-based feed-forward deep neural network that produces the emotion values of a given image. The produced emotion values are continuous values in two-dimensional space (valence and arousal), which are considered more effective than using the standard emotion categories (i.e., happiness, sadness, anger, fear, disgust, surprise and neutral) to describe emotions. [43] proposes a deep learning framework to jointly learn face representation using multimodal information. To this, the deep learning structure was composed of a set of elaborately designed CNNs and a three-layer stacked auto-encoder (SAE). The proposed system is benefitting from the complementary information in multimodal data, which achieves a higher than 99.0% recognition rate. In [44], a novel deep neural network (DNN)-driven feature learning method is proposed, which extracts from each facial image scale invariant feature transform (SIFT) features corresponding to a set of landmark points. By training the DNN model with a feature matrix consisting of the extracted SIFT feature vectors, it was possible to learn a set of optimal features that are well suitable for classifying the facial expressions across different facial views.

However, a limited number of studies specifically explored the links between QoE influencing factors (e.g., with regard to visual signal degradation) and human users' emotional responses on a behavioral level of analysis [45]. With regard to voice, [14] examines how a user's affective behavior changes with the communication quality as mediated through different network QoS conditions, and how such changes can be detected and used to predict QoE. Classification techniques based on Support Vector Machine (SVM) and k-nearest neighbors (k-NN) predict QoE with an accuracy of 67.9%. In [25], ML algorithms were trained with video recordings of participants' faces and physiological measurements to predict rated emotions in real-time. The obtained results revealed an overall good fits of the prediction models, and the performance was improved when classifying the category rather than the intensity of emotions. In [13], a QoE prediction system for video services based on emotion's analysis is provided. The major objective was to identify whether the video content watched by the viewer was in line with his/her content preferences. However, the proposed system was trained using data obtained from only 3 participants and validated by only 2 participants. In [12], the prediction of the user's QoE for video services was based on both emotional factors and network QoS parameters. Different ML algorithms were employed to test the system but the highest correlation (0.79) between subjective Mean Opinion Score (MOS) and predicted MOS was achieved with gradient based-boosting and Random Forest bagging based methods. However, also in this case, the number of video sequences (8) and testers (14) was limited for training and validation of the ML systems. Furthermore, these systems estimate the averaged MOS and not the subjective QoE perception of the single user.

In [15], [16] we have already investigated the design of an ML-based QoE prediction model for video services by considering human facial expressions and gaze direction as the training features. In [15], a crowdsourcing test was conducted where participants had to watch and rate the perceived quality of 20 videos impaired by long initial delays and re-buffering events. Three different classifiers, namely SVM with quadratic kernel, RUSBoost trees, and bagged trees, were trained with the considered features to predict the perceived QoE. Among these, the RUSBoost trees achieved the best results with the following accuracy: 44.7%, 59.4%, and 75.3% using the 5-level, 3-level, and 2-level quality scales, respectively. In [16], we conducted a laboratory test in which participants were asked to watch and rate the quality of 60 videos subject to impairments caused by blurring. In this case, the SVM with quadratic kernel classifier achieved the best results with the following accuracy: 64.5%, 78.4%, and 84.6% using the 5-level, 3-level, and 2-level quality scales, respectively. This

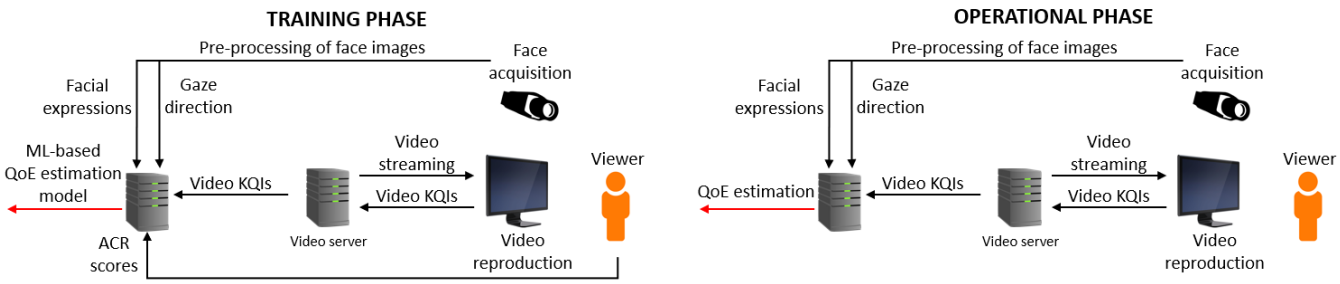


Fig. 1. Proposed methodology: training phase and operational phase.

work extends the contributions of these two papers as follows: i) we combine the subjective tests results and we extend the model combining different inputs to improve the estimation accuracy; ii) we compare the results with state-of-the-art alternative ML-based and objective QoE predictions methods.

III. Proposed Methodology

The past studies, discussed in Section II, have revealed that human emotions may be derived from facial expressions and viewer's visual attention may be gathered by gaze direction. The objective of this study is then to go deeper by looking at the relationship between this kind of viewer's information and the perceived quality. The considered scenario consists of the user consuming a video streaming service, whose face is recorded using a video camera. Observing the user's facial expressions and gaze direction, and measuring the video's KQIs during the video streaming session may be possible to estimate the degree of quality perceived. In Figure 1, we show the framework of the proposed methodology, which is composed of two phases: a training phase and an operational phase.

The training phase consists of:

1) *Data collection*: face images of the user, the video's KQIs, and feedback provided by the users (in terms of Absolute Category Rating (ACR) scores) are collected during the training phase to create a dataset.
2) *Data pre-processing*: includes the extraction of facial expression features and gaze direction from the face images of the user, and the classification of the impairment level of the measured video's KQIs. These KQIs are measured through quality metrics acquired at both end-device and server sides.
3) *Definition of the QoE model*: different QoE estimation models are defined and trained by means of different ML classifiers. The performance of these models is computed in terms of specific metrics (i.e., accuracy, specificity, sensitivity) to determine the best model in estimating the QoE based on the input features.

The output of the training phase is a ML-based QoE estimation model, which is the key part of the operational phase. The operational phase consists of:

1) *Data acquisition*: face images of the user and video's KQIs are acquired in real-time by the system.
2) *Data pre-processing*: facial expression features and gaze direction are extracted from the recorded face images; an impairment level is assigned to the video's KQIs.
3) *QoE estimation*: the QoE estimation model estimates the QoE perceived by the user based on the KQIs' impairment level and on the facial features and gaze direction extracted from the recorded face images of the user.

We want to highlight that, with regard to the facial expressions, privacy and security of users are safeguarded as the system only collects video features which do not contain any information related to the user identity. Indeed, facial expression features are immediately extracted at the face acquisition device and no pictures containing the face of the users are stored. The process is not reversible so that from the extracted features it is not possible to get back the users' face traits and to obtain any sensitive data of the user.

In the following sections we will go into the details of the training phase. Section IV describes the subjective tests we conducted for dataset generation. Section V explains how we processed the collected data before training the proposed estimation model, which is discussed in Section VI. The performance of the model is then presented in Section VII.

IV. Data Collection

For dataset generation, we conducted two experiments to collect ground-truth quality perception values with reference to video streaming services. One of these experiments, described in Section IV-A, was a crowdsourcing test in which participants had to watch and rate the quality of 20 videos impaired by long initial delays and re-buffering events. The other, described in Section IV-B, was a laboratory test in which participants were asked to watch and rate the quality of 105 videos subject to impairment caused by blurring.

A. Crowdsourcing test

The crowdsourcing test (Crowd) was carried out on the crowdsourcing platform Amazon Mechanical Turk (MTurk). We selected 5 undistorted videos from the LIVE Mobile Stall Video Database[1] [46], [47] and we created 20 test video sequences (codec: H.264, format: MP4, resolution: 1280×720 px, frame rate: 25 fps) introducing long initial delays and buffering distortions. Specifically, the KQIs considered for this

[1] http://live.ece.utexas.edu/research/LIVEStallStudy/index.html

tests are the initial delay and the number and duration of buffering events. By combining these parameters we created 4 versions of test videos:

- *Original (OR)*: 30-second version of the original video content without initial delay and buffering interruptions.
- *Long Initial (LI)*: original video content plus a long initial delay that lasted randomly in the range $8 - 20$ s.
- *Long Initial + Few Long Buffering (LIFL)*: original video content plus a long initial delay (between 8 and 20 s) plus few (between 1 and 3) long (between 10 and 15 s) buffering events.
- *Long Initial + Many Short Buffering (LIMS)*: original video content plus a long initial delay (between 8 and 20 s) plus many (between 4 and 7) short (between 2 and 4 s) buffering events.

Before conducting the test, the participant had to agree with the test conditions, which were shown on the first web page. Then, a training session was held to present to the participant an example of the impairments introduced. To this, we created 4 training videos each one with a different version of the test conditions described above with video content different from those used for the test videos. The participants had to participate to the training session before starting with the actual test. Then, the first test video to watch appeared in the next web page. The participant could decide when to start the playback of the video, during which his/her face was recorded by the PC's webcam. To facilitate the participant's attention on the video, this was shown in the center of the screen. At the end of the video reproduction, the video automatically disappeared and the participant was notified about the successful face recording and storage event. Then, the participant was allowed to rate the perceived video quality in terms of the ACR scores defined by ITU Rec. P.910, i.e., 5 (Excellent), 4 (Good), 3 (Fair), 2 (Poor), 1 (Bad) [6]. All the 20 videos were watched with the same procedure and the total time required to complete the test was around 25 minutes.

With regard to participant selection, no sex/age/country filtering was applied to the participant's requirements, but his/her HIT Approval Rate had to be greater than 90% to be accepted. This choice was aimed at increasing the probability to involve "reliable" testers. Appropriate PC requirements, screen size, display resolution, and ambient lighting were suggested in the provided test conditions but we were not able to actually verify those requirements from the platform. Therefore, we checked that these conditions were respected watching the videos of the recorded participants. We then considered only participants that watched the test videos with appropriate ambient lighting while they were sitting in front of a desk using a PC monitor. In particular, good ambient lighting was also needed to process the face images. With regard to the video playback, the test platform was set to first download the video and then play the video to avoid additional stalling events.

Many test results were discarded after the first data filtering for several reasons: i) the test was not completed, i.e., the participant did not watch and rate all the videos; ii) the participant did not record his/her face but just rated the video quality; iii) recorded videos of the participants had a bad quality and were not appropriate for extracting facial expression features. Eventually, 20 participants completed the test successfully.

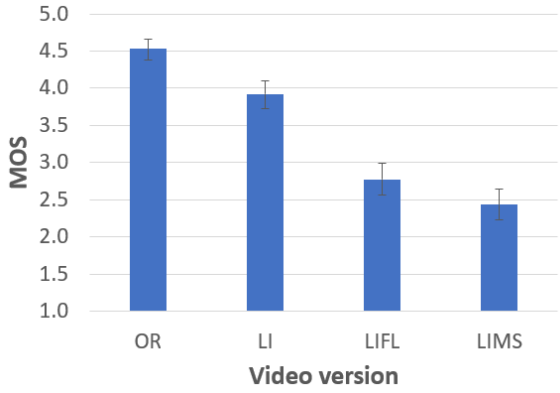


Fig. 2. Crowd test: MOS with 95% CI for each video version.

Outliers were identified through intra-rater and inter-rater reliability metrics, following the indications in [48]–[50]. The former tells to which extent the ratings of an individual rater are consistent and it was computed as the Spearman rank-order correlation coefficient (SRCC) between the participant ratings and the corresponding test conditions. The average value, computed considering all 20 participants, was 0.899. We defined as outliers the participants whose ratings obtained an intra-rater reliability lower than 0.75 [51]. We found and removed 2 outliers. After outliers removal, the average intra-rater reliability was 0.916. On the other hand, the inter-reliability describes the degree of agreement among raters and was computed as the SRCC between all participant ratings and the corresponding test conditions. The value was 1 before and after the outliers removal. Then, the final filtered data consisted of 360 ratings provided by 18 participants for the 20 test videos.

In Figure 2, we show the average MOS with 95% confidence interval (CI) computed for each test video version (OR, LI, LIFL and LIMS). The CI was computed using the Clopper-Pearson method as suggested in [52] for QoE tests characterized by a small sample size and the use of discrete bounded rating scales. As expected, the OR videos, i.e., the videos with no impairments, achieved the highest MOS. The second highest MOS is achieved by the LI videos, where only the long initial delay is present. Both these 2 video versions achieved a MOS higher than 3.5 for all video contents, which means that viewers perceived a more than sufficient video quality. Particularly, OR videos achieved good quality (MOS higher than 4). Conversely, the LIFL and LIMS videos achieved insufficient quality (MOS slightly lower than 3) and poor quality (MOS slightly lower than 2.5), respectively.

B. Laboratory test

The laboratory test (Lab) was conducted recruiting 19 healthy participants, native German speakers with normal or

corrected-to-normal vision (8 females, mean age: 29.5, range: 21–39 years). From the Sports Videos in the Wild (SVW)SVW database, 15 videos were selected which show basketball players throwing the ball inside the basket (codec: H.264, format: MP4, resolution: 480×270 px, frame rate: 30 fps) [53]. All selected videos were cut to be 20 seconds long. We created 7 versions of the 15 original videos by introducing different levels of blurring covering only the second half of the video or the entire video:

- *BLR0*: original video without blurring impairment.
- *BLR5H and BLR5E*: video post-processed with a Gaussian blurring kernel with the size of 5×5 px and standard deviation (SD) of 3 covering respectively the second half of the video and the entire video.
- *BLR10H and BLR10E*: video post-processed with a Gaussian blurring kernel with the size of 10×10 px and SD of 3 covering respectively the second half of the video and the entire video.
- *BLR15H and BLR15E*: video post-processed with a Gaussian blurring kernel with the size of 15×15 px and SD of 3 covering the second half of the video and the entire video, respectively.

The KQIs considered for this test are then the size of the blurring kernel and the blurring duration.

Before starting the test, participants received detailed written and oral instructions regarding the test procedures. Also, several training videos were shown that included all kinds of blurring impairments introduced to the test videos so as to present to the participants a preview of the video quality they had to watch and rate during the actual test. During the test the participants watched the test video sequences on an LCD computer monitor (size: 22.1×8.5×15.6 inches, resolution: 1920×1080 px, frame rate: 60 fps) and were seated at a distance of about 70 cm from the monitor. During the video playback, the face of the participant was recorded with a webcam (Logitech C920 HD Pro). After watching each video, participants had to rate the video quality on the ACR scale (same scale used for the Crowd test) and to select whether the person in the video scored a basket (this question was used to evaluate the participant's attention during the task). After 25 videos were watched and assessed, a message automatically appeared suggesting participants to have a break, which participants used to rest for 5 minutes. Then, the next group of 25 videos could be watched and rated. The time required to complete the test was approximately one hour.

Outliers were identified utilizing intra-rater and inter-rater reliability metrics as for the Crowd test (see Section IV-A). The average intra-rater reliability computed considering all 19 participants was 0.897. As for the Crowd test, we defined as outliers the participants whose ratings obtained an intra-rater reliability lower than 0.75. We found 2 outliers. After removing the 2 outliers, the average intra-rater reliability was 0.934. The inter-rater reliability value was 1 before and after the outliers removal. Then, the final filtered data consisted of 1785 ratings provided by 17 participants for the 105 test videos.

In Figure 3, we show the MOS with 95% CI computed for each video version. The CI was computed using the Clopper-Pearson method as suggested in [52]. It is evident that the size of the blurring kernel is the main factor that decreases the perceived video quality and the greater is the kernel size, the lower is the MOS. However, the QoE is also influenced by the duration of the blurring effect as video entirely covered by blurring achieved lower MOS than those partially covered (only second half of the video).

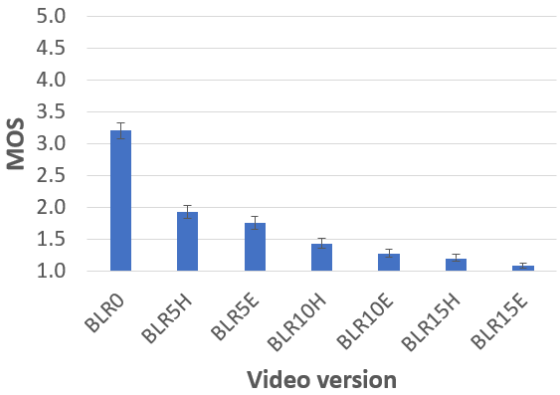


Fig. 3. Lab test: MOS with 95% CI for each video version.

V. Data Pre-processing

In this section, we describe the procedures followed to pre-process the collected data, which are illustrated in Figure 4. The final objective of the proposed data pre-processing operations is to prepare the features for the ML-based estimation model in a way that allows for merging datasets collected from different subjective tests, as it is the case of our two datasets described in Section IV. The proposed processing relies on: i) the 3 metrics that are defined in Section V-A and that make facial expression and gaze direction features independent from the duration of the recorded user's face videos; ii) the ACR and MOS normalization process described in Section V-B that allows for comparing subjective scores collected from different tests; iii) the devised impairment level feature described in Section V-C, which is aimed at making comparable the measured KQIs for streaming sessions with different types of impairments.

A. Facial expression and gaze direction features

The face acquisition device collected the images of the face of the test participants (at a rate of 30 fps) while they were watching the test videos. From these face images, we extracted facial expression and gaze direction features using the OpenFace toolkit [54]–[56]. Specifically, we based on the Facial Action Coding System (FACS) to analyze the facial expressions by means of facial Action Units (AUs) [57]. For each face image the OpenFace outputs 6 gaze direction (GD) features and 35 AUs: 18 AU_c detect the activation of a specific muscle whereas 17 AU_r detect the activation intensity (from 1 to 5)[2]. OpenFace allows for setting the face tracking confidence, which computes the confidence value of the face

[2]https://github.com/TadasBaltrusaitis/OpenFace/wiki/Output-Format

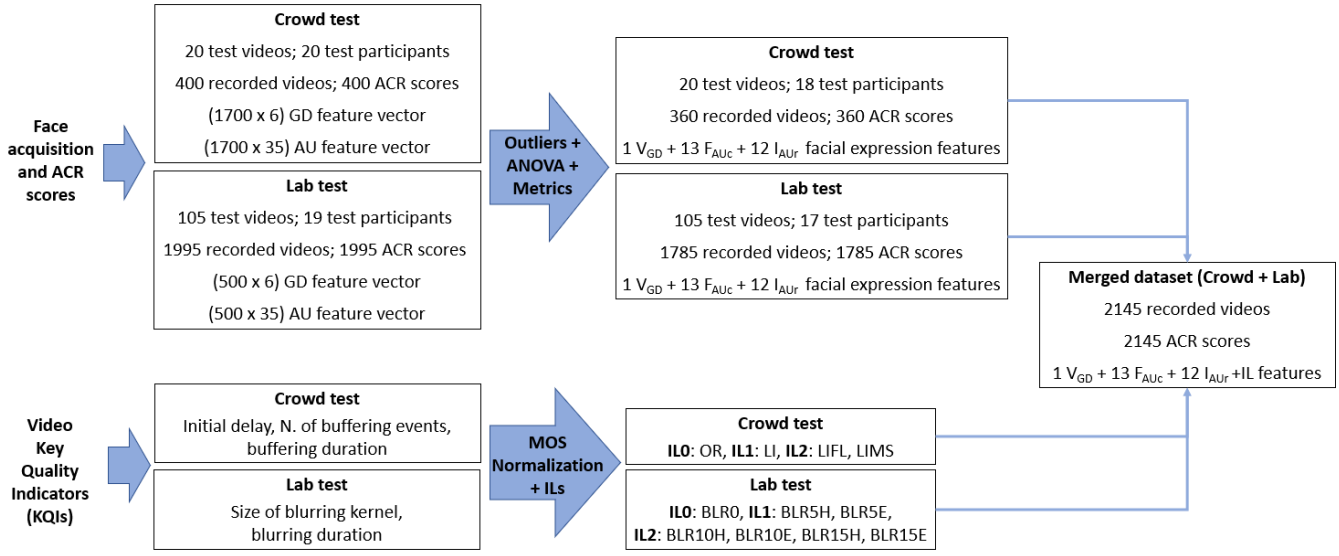


Fig. 4. Data pre-processing.

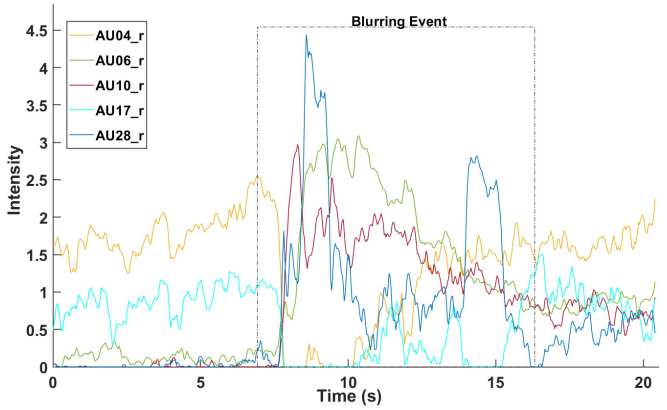


Fig. 5. Temporal evolution of the intensity of some AUs for a participant during the reproduction of the BLR10H test video.

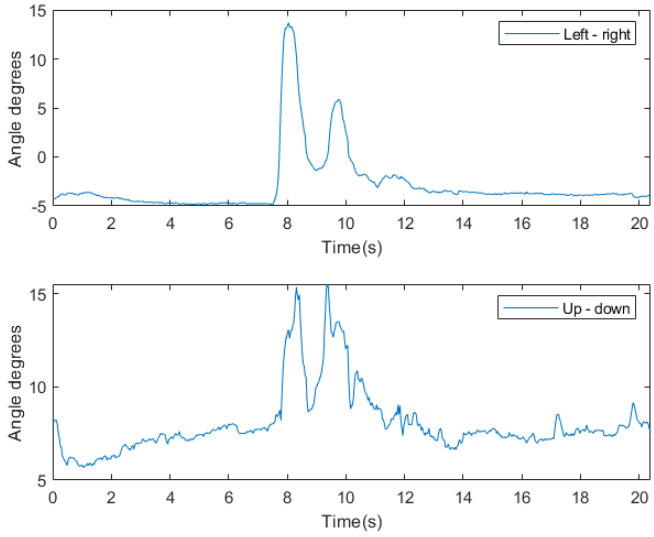


Fig. 6. Temporal evolution of the gaze direction for a participant during the reproduction of the BLR10H test video.

tracker regarding the facial landmark detection. We set a value of 98%, which means that the features were extracted only for the frames were a face was identified with a confidence higher than 98%. This guarantees high feature extraction accuracy. We also used the OpenFace option for self-calibration of the head pose tracking, which allows for automatic setting of camera parameters in terms of focal lengths and optical centre length. Figure 5 shows an example of the temporal evolution of the intensity of some AUs for a participant during the reproduction of the BLR10H test video. It can be seen as some AUs are activated more than others when the blurring starts to impair the video quality. Figure 6 shows an example of the temporal evolution of the gaze direction for a participant during the reproduction of the same test video. Again, evident variations can be noticed during the blurring event towards the left-right and up-down directions.

As summarized in Figure 4, we collected on average 1700 and 500 face images (frames from the recorded video) per participant per test video for the Crowd and Lab test, respectively. Differences are due to different test video lengths. These must be multiplied by the number of considered features, i.e., 41 (6 GD + 35 AUs) and by the number of recorded videos (400 and 1995 for the Crowd and Lab test, respectively). For the design of an effective ML algorithm, we then analysed these features to focus on those that were the most significant. For this reason, we first conducted the ANOVA analysis and then defined 3 metrics to derive a significance from the features extracted from each recorded video containing the face image of the participants: the frequency of activation of the AU_c, the intensity of activation of the AU_r, and the variance of the GD.

To try to reduce the dimensionality of the considered features, and to investigate their significance for QoE estimation, we computed the one-way ANOVA for both the Crowd and Lab tests, whose results are shown in Table II.

Note that the F value (variance between the means of the considered populations) and the p value (when $p < 0.001$ the null hypothesis is rejected) help to understand whether the considered feature is statistically relevant for QoE estimation, i.e., whether the feature training set including that feature and the feature training set not including that feature are significantly different.

Note that we are providing in this table only the values for the features which were significant for both the Crowd and Lab test. With regard to the eye gaze features, the x and y direction vectors for both the leftmost (eye0) and rightmost (eye1) eyes are statistically relevant, as well as the x and y gaze direction angles in radians. With regard to the AUs, both the intensity and activation of inner brow raiser (AU01), outer brow raiser (AU02), brow lowerer (AU04), nose wrinkler (AU09), lip corner depressor (AU15), chin raiser (AU17), and lip tightener (AU23) are statistically relevant. The intensity of lid tightener (AU07), lip corner puller (AU12), dimpler (AU14), lip stretcher (AU20), lips part (AU25), and jaw drop (AU26), and the activation of upper lid raiser (AU05), cheek raiser (AU06), upper lip raiser (AU10), lip suck (AU28), and blink (AU45) are also significant. These results highlight that not all the features are significant for both tests but different impairments can provide different reactions in terms of perceived quality. Moreover, the AUs can provide insights regarding the emotions felt by the viewers [58]. For example, AU23 must be present for anger, AU9 and AU10 for disgust, AU12 for happiness, either AU01+AU02 or AU05 for surprise and either AU01+AU04+AU15 or AU11 for sadness.

On the basis of this ANOVA analysis, the 6 GD and 25 AU features in Table II are the most significant for QoE estimation and we computed 3 metrics from these feature arrays as explained in the following. We computed the *frequency of activation* $F_{AU_c^j}$ of the j-th AU_c over the activation of all the AU_c, as in Eq. (1). Activation means that the AU_c assumes a value greater than 0.

$$F_{AU_c^j} = \frac{\sum_{n=1}^{N} a_n^j}{\sum_{n=1}^{N}\sum_{j=1}^{J} a_n^j} \tag{1}$$

where $a_n^j = 1$ if $AU_{c,n}^j > 0$, otherwise $a_n^j = 0$, $n \in \{1, 2, ..., N\}$ identifies the video frames, and $j \in \{1, 2, ..., J\}$ identifies the considered AU_c ($J = 18$).

Furthermore, we computed the *intensity of the activated* AU_r, $I_{AU_r^k}$, as in Eq. (2).

$$I_{AU_r^k} = \sum_{n=1}^{N} AU_{r,n}^k \tag{2}$$

where $n \in \{1, 2, ..., N\}$ identifies the video frames, and $k \in \{1, 2, ..., K\}$ identifies the considered AU_r ($K = 17$).

Finally, we define the *variance of the GD*, V_{GD^g}, which is computed as the variance of the position of the eyes' pupils over the whole sequence of frames of the viewer face video, as in Eq. (3).

$$V_{GD^g} = \frac{\sum_{n=1}^{N}(GD_n^g - \overline{GD^g})}{N} \tag{3}$$

TABLE II
ANOVA RESULTS FOR THE CONSIDERED FEATURES.

Feature	Crowd		Lab	
	F	p	F	p
Eye0-gaze x vector	8.35E+14	< 0.001	4.06E+12	< 0.001
Eye0-gaze y vector	8.09E+14	< 0.001	4.56E+12	< 0.001
Eye1-gaze x vector	6.61E+14	< 0.001	5.67E+12	< 0.001
Eye1-gaze y vector	6.61E+14	< 0.001	5.03E+12	< 0.001
Gaze x angle	8.00E+14	< 0.001	5.46E+12	< 0.001
Gaze y angle	1.93E+14	< 0.001	4.72E+12	< 0.001
$AU01_c$	1.49E+14	< 0.001	1.50E+14	< 0.001
$AU02_c$	-1.52E+13	$> \mathbf{0.001}$	1.84E+14	< 0.001
$AU04_c$	2.16E+14	< 0.001	7.95E-01	$> \mathbf{0.001}$
$AU07_c$	1.32E+14	< 0.001	1.10E+14	$> \mathbf{0.001}$
$AU09_c$	1.57E+14	< 0.001	1.00E+14	$> \mathbf{0.001}$
$AU12_c$	1.05E+14	$> \mathbf{0.001}$	1.28E+14	< 0.001
$AU14_c$	2.30E+14	< 0.001	1.32E+14	< 0.001
$AU15_c$	1.41E+14	< 0.001	1.08E+14	$> \mathbf{0.001}$
$AU17_c$	1.36E+14	< 0.001	2.70E+14	< 0.001
$AU20_c$	6.33E+14	< 0.001	1.26E+14	< 0.001
$AU23_c$	1.65E+14	< 0.001	1.46E+14	< 0.001
$AU25_c$	6.27E+14	< 0.001	1.42E+14	< 0.001
$AU26_c$	2.09E+14	< 0.001	1.09E+14	$> \mathbf{0.001}$
$AU01_r$	-1.50E+13	$> \mathbf{0.001}$	4.10E+14	< 0.001
$AU02_r$	1.44E+14	< 0.001	9.79E+04	< 0.001
$AU04_r$	1.32E+14	< 0.001	9.79E-01	$> \mathbf{0.001}$
$AU05_r$	1.33E+14	< 0.001	1.01E+14	$> \mathbf{0.001}$
$AU06_r$	1.30E+14	< 0.001	9.70E-01	$> \mathbf{0.001}$
$AU09_r$	1.31E+14	< 0.001	9.84E-01	$> \mathbf{0.001}$
$AU10_r$	1.40E+14	< 0.001	9.59E+04	< 0.001
$AU15_r$	1.42E+14	< 0.001	9.81E-01	$> \mathbf{0.001}$
$AU17_r$	1.35E+14	< 0.001	9.40E-01	$> \mathbf{0.001}$
$AU23_r$	1.23E+14	< 0.001	9.87E-01	$> \mathbf{0.001}$
$AU28_r$	1.26E+14	< 0.001	9.62E-01	$> \mathbf{0.001}$
$AU45_r$	1.34E+14	< 0.001	9.99E-01	$> \mathbf{0.001}$

where $g \in \{1, 2, 3, 4, 5, 6\}$ identifies respectively the six single features that determine the gaze direction, i.e., $0x$, $0y$, $1x$, $1y$, $Angle_x$, and $Angle_y$. From these, we were able to estimate which part of the video the visual attention of the viewer was focused on and, most important, to notice whether the viewer look the other way. This was totally feasible for the Lab test as it was a controlled test during which all test conditions were verified to avoid participant's defections. On the contrary, the Crowd test was more difficult to control. However, as participants were recorded, we verified that all participants watched the videos while they were sitting in front of a desk using a PC monitor.

Therefore, the facial expression features selected and used to train the ML-based QoE estimation model presented in Section VI-C are: 1 V_{GD}, 13 F_{AU_c} and 12 I_{AU_r}. It is important to highlight that by computing these metrics we obtain the same number of features for each recorded video, regardless the video length. This is important to make feature datasets compatible for merging.

B. ACR scores

As illustrated in Figure 4 and discussed in Sections IV-A and IV-B, we collected 400 and 1995 ACR scores for the Crowd and Lab test, respectively. After outliers removal, we kept a total of 360 and 1785 ACR scores, respectively.

As it can be noticed in Figures 2 and 3, participants of different tests used the MOS scale differently for rating the

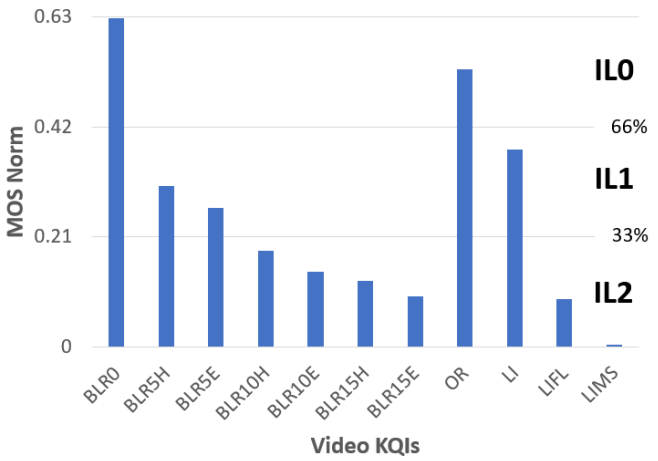


Fig. 7. Normalized MOS computed for both the datasets and definition of the Impairment Levels (ILs).

video quality. Indeed, blurring impairments were perceived as more annoying than buffering events. The greatest MOS for the Lab test is slightly greater than 3, whereas for the Crowd test the MOS ranges from 2.5 to 4.5. To be able to merge these datasets, we normalized the ACR scores applying for both the datasets the mean normalization method as follows:

$$ACR_{norm} = \frac{ACR - ACR_{mean}}{ACR_{max} - ACR_{min}} \quad (4)$$

where $ACR_{mean} = MOS$, $ACR_{max} = 5$ and $ACR_{min} = 1$. ACR_{norm} values are mapped in the [-1;1] range. We then computed the normalized MOS from the obtained normalized ACR scores, which is shown in Figure 7 (for easier comparison, MOS values were also shifted to set the lowest value to 0), from which we define the impairment levels as described in the next section. We then merged the ACR score datasets together (and corresponding features) for training the ML-based QoE estimation model presented in Section VI-C.

C. Impairment level of video KQIs

As already mentioned, KQIs for the considered applications are also considered as an additional input for the QoE estimation. The video KQIs considered for the Crowd test are the initial delay, the number of buffering events and the buffering duration, whereas those considered for the Lab test are the size of the blurring kernel and the blurring duration. Besides the facial metrics introduced in Section V-A, we aimed to consider for training the ML-based model also a feature that contains information regarding these KQIs related to the different test condition.

Due to the differences in the KQIs characterizing different service settings, we define the *Impairment level* (IL) feature which could be used for any settings. It has the objective to highlight the level of impairment introduced to the video through the specific setting measured KQIs. To define the IL we relied on the normalized MOS shown in Figure 7 and we decided to consider 3 ILs. We considered the worst test conditions (LIMS) as the zero quality and the best test condition (BLR0) as the 100% quality. IL0 includes test conditions achieving a MOS between 67% and 100% of the maximum normalized MOS. IL1 includes test conditions achieving a MOS between 33% and 66% of the maximum quality, and finally IL2 considers test conditions achieving MOS lower than 33% of the maximum quality. This approach discretizes the measured KQIs and allows for comparing them even if they are not typically comparable as identify different impairments. We based on this approach assuming that when users perceive lower quality than expected, their disappointment is reflected in their facial expressions and the higher is the disappointment the different is the facial expressions' variation.

The IL is then an additional feature that we considered for training the ML-based model. We identify this feature as KQI impairment level $KQIIL_l$ where $l \in \{0, 1, 2\}$ identifies the IL. With this feature is then possible to merge datasets obtained from tests considering different KQIs.

VI. ML-based QoE Model

In this section, we describe the considered features, (Section V-A), the considered data augmentation approach (Section VI-B) and the proposed QoE estimation models (Section VI-C).

A. Features

When training a ML-based model two types of variable must be considered: independent and dependent variables. The former are the controlled inputs and are typically referred to as features. The latter represent the output or outcome resulting from altering these inputs.

In our case, the considered features are 27: 1 V_{GD}, 13 F_{AU_c}, 12 I_{AU_r} and 1 $KQIIL_l$. The outputs are the 2145 ACR scores. The feature matrix is then 2145×27. These features are the result of the data pre-processing operations described in Section V and illustrated in Figure 4.

B. Data Augmentation

Since the obtained dataset had a class imbalanced problem (number of ACR scores: '1': 970, '2': 593, '3': 326, '4': 150, '5': 106), we considered a class over-sampling method for creating synthetic samples instead of replicating samples in the dataset. Specifically, the adaptive synthetic (ADASYN) algorithm which balances binary classes was used for the present study [59]. The ADASYN algorithm aims to improve class balance by synthetically creating new examples from the minority class via linear interpolation between existing minority class examples. The idea behind this algorithm is using a weighted data distribution for different minority class samples with respect to their capacity in learning [60]. The algorithm works creating more examples in the vicinity of the boundary between the two classes than in the interior of the minority class. In our case, the dataset has been extended to also handle multiple classes (k classes) by simply calling the ADASYN algorithm for $k-1$ classes that are not the majority class and merging the obtained results. After the DA process, the number of ACR scores for each class was: '1': 970, '2': 893, '3': 626, '4': 450, '5': 406, for a total of 3345 values.

TABLE III
THE THREE CONSIDERED QUALITY SCALES.

Rate	ACR 5-level quality scale	3-level quality scale	2-level quality scale
1	Excellent	Very good	Very good
2	Good		
3	Fair	Sufficient	
4	Poor	Bad	Bad
5	Bad		

As explained above, the majority class has always the same dimension. The other classes were augmented to reduce the gap between the classes.

C. QoE estimation model

We defined two ML-based QoE estimation models based on the features described in Section VI-A. Specifically, the QoE models are as follows:

- *AU&GDtoQoE*: takes as input for each video the features related to the viewer's facial expressions and gaze direction (V_{GD}, F_{AU_c}, and I_{AU_r}), and as output the viewer's QoE.
- *AU&GD&KQItoQoE*: takes as input for each video the features related to the viewer's facial expressions and gaze direction (V_{GD}, F_{AU_c}, and I_{AU_r}), and the impairment level of the measured video KQIs ($KQIIL_l$) and as output the viewer's QoE.

Both the two QoE estimation models were implemented with two different classifiers, namely an SVM with a quadratic kernel and a k-NN, to find a pattern within the features dataset that could describe a correlation with the QoE final score provided by the viewers. The k-NN has been set using the City-block distance metric, considering the neighbours' number equal to 1 and the Square inverse distance weight. The SVM has been set several times by changing the box constraint level and the kernel scale size but it always achieved lower results than the k-NN. The classifiers were implemented with the MATLAB software using specific machine learning libraries and the parallel computing toolbox supported by the CUDA drivers[3]. Finally, for QoE prediction we considered the three quality scales in Table III. The best performance results were obtained with the k-NN classifier in all cases.

VII. EXPERIMENT RESULTS

In this section, we evaluate the performance of the ML-based QoE estimation models proposed in Section VI-C using three performance metrics for ML, i.e., accuracy, sensitivity, and specificity.

Firstly, we compare the model performance for different training/validation rates (90%/10%, 80%/20%, and 70%/30%) and k-fold cross-validation ($k = 5$, and $k = 10$). For this comparison we used the AU&GD&KQItoQoE model implemented with the k-NN classifier using the original Crowd+Lab dataset without data augmentation. Results are provided in Table IV. The greatest accuracy, 87.8% and 87.7%, is achieved respectively by the 70%/30% combination with $k = 5$ and by the 80%/20% combination with $k = 10$. Then, we analyse the specificity and sensitivity metrics. The specificity is comparable for both the combinations whereas the sensitivity achieves higher values for the 70%/30% combination with $k = 5$, in particular for the classes '3' and '5'. We then consider this combination to compare the performance of the QoE model trained with (AU&GD&KQItoQoE) and without (AU&GDtoQoE) the $KQIIL_l$ feature. Also in this case the models were implemented with the k-NN classifier using the original Crowd+Lab dataset without data augmentation. Results are shown in Table V. It can be seen that considering the $KQIIL_l$ feature the accuracy increases from 81.9% to 86.8%. Specificity and sensitivity are increased as well when this feature is considered for model training. Furthermore, using the aforementioned setting, we also trained the model without the GD feature. However, results were quite poor as accuracy reach 60%, specificity ranges from 70% to 90% and sensitivity ranges from 20% to 60% for all the 5 classes. This confirms the results achieved in [15], [16] where the GD feature allowed to achieve higher QoE estimation accuracy.

Finally, in Table VI we compare the accuracy, sensitivity, and specificity metrics obtained with the AU&GDtoQoE and AU&GD&KQItoQoE models implemented with the k-NN classifier using the augmented datasets. In particular, for each model we performed two cross-dataset evaluations (applying data augmentation on the training dataset) and an evaluation using the combined (Crowd+Lab) augmented dataset (see Section VI-B) and a 70%/30% training/validation combination with $k = 5$ (best combination obtained in Table IV). First of all, it can be noticed that the k-NN classifier performs better when trained with all features (AU, GD and $KQIIL_l$) in terms of all the considered metrics and for each of the 3 quality scales, which confirm the results of Table V. This suggests that the impairment level defined in Section V-C is a feature that allows the ML-based model to better define a pattern within the features dataset that describes a correlation with the ACR scores. Both the cross-dataset evaluations achieved good performance but the greatest performance are achieved training on the combined augmented datasets. For this case, an accuracy of 93.9%, 97.1% and 98.1% is achieved for the 5-level, 3-level and 2-level quality scales. Indeed, in general the performance achieved using the 2-level quality scale are higher than those achieved using the 3-level quality scale, which in turn are higher than those achieved using the 5-level quality scale. However, considering that using the 5-level quality scale we are defining a mapping between the complete ACR quality scale and the viewer's perceived video quality derived from his/her facial expressions and gaze direction, the achieved performance remain relevant. Nonetheless, the greater performance achieved by the 3-level and 2-level quality scales may be due to the fact that it is easier to identify extreme emotions (such as positive or negative) from facial expressions whereas halfway emotions are more difficult to be identified. Still, the use of the 3-level and 2-level quality scales may be useful for QoE estimation as it may help to identify when the viewer is not satisfied by the perceived quality, which may be due to the distortions affecting the video

[3]https://docs.nvidia.com/cuda

TABLE IV
COMPARISON OF ACCURACY, SENSITIVITY, AND SPECIFICITY METRICS FOR DIFFERENT TRAINING/VALIDATION RATES AND K-FOLD CROSS-VALIDATION. RESULTS CONCERN THE AU&GD&KQITOQOE MODEL IMPLEMENTED WITH THE K-NN CLASSIFIER USING THE ORIGINAL CROWD+LAB DATASET WITHOUT DATA AUGMENTATION.

Training	Validation	k	5-level quality scale		
			Specificity	Sensitivity	Accur.
90%	10%	5	$1 = 91.73\%$ $2 = 95.83\%$ $3 = 97.59\%$ $4 = 98.32\%$ $5 = 98.19\%$	$1 = 94.79\%$ $2 = 87.23\%$ $3 = 77.67\%$ $4 = 72.92\%$ $5 = 72.22\%$	86.9%
80%	20%	5	$1 = 91.54\%$ $2 = 96.11\%$ $3 = 97.03\%$ $4 = 98.25\%$ $5 = 98.37\%$	$1 = 95.40\%$ $2 = 87.02\%$ $3 = 75.50\%$ $4 = 72.75\%$ $5 = 70.10\%$	86.7%
70%	30%	5	$1 = 91.43\%$ $2 = 96.20\%$ $3 = 97.69\%$ $4 = 98.44\%$ $5 = 98.69\%$	$1 = 95.98\%$ $2 = 88.57\%$ $3 = 87.18\%$ $4 = 74.40\%$ $5 = 73.91\%$	87.8%
90%	10%	10	$1 = 91.89\%$ $2 = 95.72\%$ $3 = 95.79\%$ $4 = 99.32\%$ $5 = 98.53\%$	$1 = 95.80\%$ $2 = 84.74\%$ $3 = 87.63\%$ $4 = 77.26\%$ $5 = 78.67\%$	86.8%
80%	20%	10	$1 = 93.82\%$ $2 = 96.11\%$ $3 = 97.21\%$ $4 = 97.51\%$ $5 = 98.60\%$	$1 = 95.53\%$ $2 = 87.26\%$ $3 = 78.67\%$ $4 = 78.57\%$ $5 = 66.67\%$	87.7%
70%	30%	10	$1 = 91.33\%$ $2 = 95.82\%$ $3 = 96.80\%$ $4 = 98.01\%$ $5 = 98.48\%$	$1 = 93.68\%$ $2 = 86.19\%$ $3 = 81.64\%$ $4 = 65.29\%$ $5 = 58.69\%$	86.2%

TABLE V
COMPARISON OF ACCURACY, SENSITIVITY, AND SPECIFICITY METRICS OBTAINED WITH THE AU&GDTOQOE AND AU&GD&KQITOQOE MODELS IMPLEMENTED WITH THE K-NN CLASSIFIER USING THE ORIGINAL CROWD+LAB DATASET WITHOUT DATA AUGMENTATION.

Model	5-level quality scale		
	Specificity	Sensitivity	Accuracy
AU&GDtoQoE	$1 = 90.49\%$ $2 = 93.19\%$ $3 = 94.63\%$ $4 = 98.09\%$ $5 = 98.20\%$	$1 = 87.52\%$ $2 = 84.66\%$ $3 = 82.87\%$ $4 = 70.65\%$ $5 = 70.97\%$	81.9%
AU&GD&KQItoQoE	$1 = 92.72\%$ $2 = 94.73\%$ $3 = 96.41\%$ $4 = 98.32\%$ $5 = 99.26\%$	$1 = 92.83\%$ $2 = 85.21\%$ $3 = 83.50\%$ $4 = 78.75\%$ $5 = 79.10\%$	86.8%

reproduction. For example, a 3-level quality scale is used in [4] to evaluate whether the video quality is acceptable or annoying (Not Annoying / Annoying but acceptable / Not acceptable).

A. Comparison with state-of-the-art QoE models

In this section we compare the performance of the proposed ML-based QoE estimation model with those of state-of-the-art QoE models in terms of the Pearson correlation coefficient (PCC) and Root Mean Square Error (RMSE) computed between the actual MOS and the estimated MOS. To this, we trained a Coarse Tree regression model using the MOS values as independent variable (and not the ACR as in the previous cases) and considering the 5-level quality scale (i.e., the MOS scale). The regression model uses the Mean Squared Error (MSE) as splitting criterion and it has the minimum leaf size set to 36. The following results have been obtained using the 5 fold cross validation with a 70%/30% training/validation combination. The achieved PCC and RMSE are respectively 0.989 and 0.11.

To the best of the authors' knowledge, the only valuable ML-based model based on facial expressions for QoE estimation is [12]. Indeed, the model in [13] was trained and validated with data from only 3 and 2 subjects, respectively, which is not statistically relevant. Besides facial expression features, the model in [12] was trained using also network QoS parameters. The testers involved for creating the dataset for this model were 16 but no details are given regarding the number of videos each tester had to watch and evaluate. This model, with its dataset, achieved a PCC of 0.79, lower than the PCC achieved by our model, and an RMSE of 0.55, which is five times the RMSE achieved by our model.

Then, we compare the performance of our model with those of QoE objective models for the single tests. With regard to the Lab test, we used the NR perceptual blur metric defined in [61] and implemented with MATLAB [62] to measure the quality of the 105 test videos used for the Lab test. This metric basically compares the intensity variations between the neighboring pixels of the original image and the neighboring pixels of the same image after a low-pass filtering step. This metric achieved a PCC of 0.81 and an RMSE of 0.35. The PCC is lower than that of our model and the RMSE is three times greater, even considering that our model achieved that performance for the entire dataset (Lab+Crowd). Note that to compute properly the RMSE, we normalized the MOS values to the $0-1$ scale used by the blur metric to evaluate the image quality.

With regard to the Crowd test, we used the ITU-T P.1203 model [63] implemented with Python [51], [64][4] to measure the quality of the 20 test videos used for the Crowd test. The P.1203 framework includes a video quality estimation module, an audio quality estimation module, and a quality integration module. We set the model to the *Mode 3* to access the full bitstream and we considered the main output *O.46*, i.e., the final media session quality score. This model achieved a PCC of 0.987 and an RMSE of 0.28. While the PCC is comparable with that of our model, the RMSE is more than doubled. In particular, the higher RMSE is due by an overestimation of the MOS predicted by the P.1203 model with respect to the actual MOS. The reason may be that this model is solely based on objective measures of the video quality whereas the proposed model also considers subjective factors derived from facial expressions that support the performance improvement. Again, it must be considered that our model achieved that performance for the entire dataset (Lab+Crowd).

[4]https://github.com/itu-p1203/itu-p1203

TABLE VI
COMPARISON OF ACCURACY, SENSITIVITY, AND SPECIFICITY METRICS OBTAINED WITH THE AU&GDTOQOE AND AU&GD&KQITOQOE MODELS IMPLEMENTED WITH THE K-NN CLASSIFIER USING THE AUGMENTED DATASETS.

Model	Training dataset	Validation dataset	5-level quality scale			3-level quality scale			2-level quality scale		
			Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy
AU&GDtoQoE	Crowd augmented	Lab	$1 = 98.00\%$ $2 = 96.39\%$ $3 = 88.43\%$ $4 = 95.41\%$ $5 = 92.77\%$	$1 = 74.19\%$ $2 = 79.17\%$ $3 = 82.61\%$ $4 = 83.33\%$ $5 = 76.74\%$	78.0%	$1/2 = 95.72\%$ $3 = 93.20\%$ $4/5 = 85.77\%$	$1/2 = 80.56\%$ $3 = 82.16\%$ $4/5 = 87.70\%$	84.4%	$1/2 = 95.36\%$ $3/4/5 = 95.36\%$	$1/2 = 91.39\%$ $3/4/5 = 91.39\%$	91.4%
	Lab augmented	Crowd	$1 = 88.44\%$ $2 = 89.19\%$ $3 = 96.94\%$ $4 = 99.22\%$ $5 = 99.90\%$	$1 = 87.71\%$ $2 = 83.01\%$ $3 = 81.77\%$ $4 = 74.10\%$ $5 = 77.78\%$	83.5%	$1/2 = 71.50\%$ $3 = 98.76\%$ $4/5 = 99.53\%$	$1/2 = 98.45\%$ $3 = 94.07\%$ $4/5 = 89.82\%$	93.7%	$1/2 = 91.17\%$ $3/4/5 = 91.17\%$	$1/2 = 97.84\%$ $3/4/5 = 97.84\%$	95.6%
	70% Crowd+Lab augmented $k = 5$	30% Crowd+Lab augmented $k = 5$	$1 = 91.43\%$ $2 = 96.20\%$ $3 = 97.69\%$ $4 = 98.44\%$ $5 = 98.69\%$	$1 = 95.98\%$ $2 = 86.56\%$ $3 = 80.18\%$ $4 = 79.40\%$ $5 = 78.91\%$	87.8%	$1/2 = 73.33\%$ $3 = 98.50\%$ $4/5 = 99.41\%$	$1/2 = 98.15\%$ $3 = 93.08\%$ $4/5 = 88.86\%$	93.6%	$1/2 = 94.85\%$ $3/4/5 = 94.85\%$	$1/2 = 99.02\%$ $3/4/5 = 99.02\%$	96.8%
AU&GD&KQItoQoE	Crowd augmented	Lab	$1 = 99.99\%$ $2 = 98.92\%$ $3 = 87.17\%$ $4 = 95.11\%$ $5 = 98.49\%$	$1 = 88.57\%$ $2 = 77.60\%$ $3 = 93.75\%$ $4 = 87.62\%$ $5 = 77.23\%$	85.2%	$1/2 = 97.63\%$ $3 = 95.25\%$ $4/5 = 91.77\%$	$1/2 = 87.23\%$ $3 = 89.78\%$ $4/5 = 92.54\%$	90.5%	$1/2 = 98.13\%$ $3/4/5 = 98.13\%$	$1/2 = 88.09\%$ $3/4/5 = 88.09\%$	95.9%
	Lab augmented	Crowd	$1 = 88.02\%$ $2 = 91.50\%$ $3 = 98.09\%$ $4 = 99.52\%$ $5 = 99.81\%$	$1 = 91.76\%$ $2 = 88.03\%$ $3 = 79.80\%$ $4 = 81.05\%$ $5 = 87.50\%$	86.5%	$1/2 = 98.41\%$ $3 = 97.75\%$ $4/5 = 99.99\%$	$1/2 = 98.64\%$ $3 = 97.78\%$ $4/5 = 93.33\%$	97.7%	$1/2 = 97.18\%$ $3/4/5 = 97.18\%$	$1/2 = 99.00\%$ $3/4/5 = 99.00\%$	99.0%
	70% Crowd+Lab augmented $k = 5$	30% Crowd+Lab augmented $k = 5$	$1 = 95.85\%$ $2 = 96.18\%$ $3 = 98.41\%$ $4 = 98.61\%$ $5 = 99.29\%$	$1 = 96.87\%$ $2 = 93.23\%$ $3 = 90.26\%$ $4 = 89.47\%$ $5 = 89.67\%$	93.9%	$1/2 = 92.17\%$ $3 = 97.89\%$ $4/5 = 99.35\%$	$1/2 = 98.38\%$ $3 = 96.08\%$ $4/5 = 94.02\%$	97.1%	$1/2 = 91.33\%$ $3/4/5 = 91.33\%$	$1/2 = 99.36\%$ $3/4/5 = 99.36\%$	98.1%

B. Impact of face acquisition frequency and computational cost

The proposed system is based on the acquisition of the user's face and extraction of the relevant features, which have to be performed at the user side, possible in real-time, as described in Section III. However, in this paper, these procedures have been implemented offline in our lab server and we focused on the design of the system with particular attention to the QoE estimation performance, without considering the engineering issues related to the implementation of the estimator in the user terminal. However, in this section, we discuss the impact of the face acquisition frequency on the QoE estimation as well as the computational cost of the procedures for the extraction of facial expression features and gaze direction from the acquired face image. To this, we conducted specific experiments using the Lab dataset.

We define the *sample rate* as the rate at which the user's face video is sampled to obtain the frames that are used to extract the features needed to build the dataset. We have evaluated its impact on the system performance by considering 6 sample rates: 1 sample every h frames, with $h = 5, 10, 15, 20, 25, 30$. The aim is to investigate whether the QoE estimation accuracy would decrease in case the sample rate decreases too much with respect to the ideal case (i.e., all frames are used for QoE estimation). We obtained 6 feature matrices (as defined in Section VI-A but without the KQI feature), one for each of the 6 considered sample rates. We then computed the PCC between these feature matrices and the ideal (the one without sampling) feature matrix. Table VII reports PCC results. It can be seen as the PCC achieves values higher than 0.94 even when only 1 face image every 30 frames is used to extract the features to create the training dataset. This means that training the model with this reduced dataset would not impact on the achieved estimation accuracy.

With regard to the computational cost of the acquisition device, we measured the system performance during the feature extraction. The average time needed to extract facial expression features and gaze direction from a 20-second long video with 30 fps is 19.7 s. When 30 fps videos are considered, on average the duration of the video is needed to extract the features. However, by decreasing the sample rate the extraction time will also decrease. As an example, the extraction time needed by considering a sample rate of 1/5 is 3.5 s. The tested system was run on a laptop with 16 GB of RAM and CPU Intel Core I7-7700 HQ. The feature extraction process required, on average, 580 MB of RAM and the 47% of the CPU.

TABLE VII
IMPACT OF FACE ACQUISITION FREQUENCY.

Sample rate	PCC
1/5	0.995
1/10	0.985
1/15	0.975
1/20	0.964
1/25	0.956
1/30	0.945

VIII. CONCLUSION

In this paper, we focused on an ML-based approach to define a QoE estimation model able to accurately estimate the QoE perceived by users watching video sequences. The estimation was solely based on the analysis of the facial expressions (in terms of AUs) and gaze directions of the user while watching the video and on the level of quality impairment deduced by the measured video KQIs (by defining the KQI-related feature named KQIIL). Two datasets were used for training and validation. Among the different classifiers used to train the model, the k-NN classifier achieved the best performance in terms of specificity, sensitivity and accuracy metrics. Moreover, the model performed better when trained

with all the features (AU, gaze and KQIIL) than with only the AU and gaze features, underlining that the information about KQIIL is directly connected to the perceived quality and allows for achieving better QoE estimation performance.

The proposed model outperforms state-of-the-art ML-based and objective video QoE models in terms of the aforementioned metrics as well as in terms of PCC and RMSE. Specifically, we achieved performance in terms of PCC and RMSE of 0.989 and 0.11, respectively. State-of-the-art studies that consider facial expression estimation were able to achieve a PCC of 0.79, which is quite lower than what we achieved, and an RMSE of 0.55, which is five times the one we obtained. We also compared our results with two relevant QoE objective models. One that is appropriate for the blurring impairment with introduced in one dataset, which achieved a PCC of 0.81 and an RMSE of 0.35, which are again worse than ours. Another that is optimal for measuring impairments typical of video streaming scenarios that we considered in the second dataset of our experiments, i.e., the P.1203 framework. The latter achieved a PCC of 0.987 and an RMSE of 0.28. While the PCC is comparable with that of our model, the RMSE is more than double with respect to ours.

In future works, we aim to apply the proposed methodology on different applications, such as web browsing and VoIP. Indeed, it is highly likely that subjective factors, such as user's facial expressions and gaze direction, will also provide relevant insights regarding the perceived quality when different multimedia services are considered. Also, we aim to collect additional datasets of features which may help to improve the QoE estimation performance of the model. Additional features may concern, for example, context information regarding the environment (e.g., lighting, external noise), the social factor (if a service is consumed alone or with other people) and device information (e.g., fixed or mobile device, screen size and resolution).

REFERENCES

[1] Cisco. (2020) Cisco visual networking index: Forecast and trends, 2017–2022 white paper. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html

[2] L. Skorin-Kapov, M. Varela, T. Houndefinedfeld, and K.-T. Chen, "A Survey of Emerging Concepts and Challenges for QoE Management of Multimedia Services," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 2s, May 2018.

[3] A. Ahmad and L. Atzori, "MNO-OTT Collaborative Video Streaming in 5G: The Zero-rated QoE Approach for Quality and Resource Management," *IEEE Trans. on Network and Service Management*, vol. 17, no. 1, pp. 361–374, 2020.

[4] J. Li, L. Krasula, Y. Baveye, Z. Li, and P. Le Callet, "AccAnn: A New Subjective Assessment Methodology for Measuring Acceptability and Annoyance of Quality of Experience," *IEEE Trans. on Multimedia*, vol. 21, no. 10, pp. 2589–2602, Oct 2019.

[5] P. Le Callet, S. Möller, and A. Perkis. (2012) Qualinet White Paper on Definitions of Quality of Experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, Version 1.2, March 2013.

[6] ITU, "Subjective video quality assessment methods for multimedia applications." Recommendation ITU-T P.910, 2008.

[7] P. Casas, A. D'Alconzo, P. Fiadino, A. Bär, A. Finamore, and T. Zseby, "When YouTube Does not Work—Analysis of QoE-Relevant Degradation in Google CDN Traffic," *IEEE Trans. on Network and Service Management*, vol. 11, no. 4, pp. 441–457, Dec 2014.

[8] S. Kumar, R. Devaraj, A. Sarkar, and A. Sur, "Client-Side QoE Management for SVC Video Streaming: An FSM Supported Design Approach," *IEEE Trans. on Network and Service Management*, vol. 16, no. 3, pp. 1113–1126, Sep. 2019.

[9] P. Casas, M. Seufert, F. Wamser, B. Gardlo, A. Sackl, and R. Schatz, "Next to You: Monitoring Quality of Experience in Cellular Networks From the End-Devices," *IEEE Trans. on Network and Service Management*, vol. 13, no. 2, pp. 181–196, June 2016.

[10] H. Yan, T. Lin, C. Gao, Y. Li, and D. Jin, "On the Understanding of Video Streaming Viewing Behaviors Across Different Content Providers," *IEEE Trans. on Network and Service Management*, vol. 15, no. 1, pp. 444–457, March 2018.

[11] U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J. Antons, K. Y. Chan, N. Ramzan, and K. Brunnström, "Psychophysiology-Based QoE Assessment: A Survey," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 6–21, 2017.

[12] L. Amour, M. I. Boulabiar, S. Souihi, and A. Mellouk, "An improved QoE estimation method based on QoS and affective computing," in *2018 Int. Symposium on Programming and Systems (ISPS)*, 2018, pp. 1–6.

[13] X. Tao, L. Dong, Y. Li, J. Zhou, N. Ge, and J. Lu, "Real-time personalized content catering via viewer sentiment feedback: a QoE perspective," *IEEE Network*, vol. 29, no. 6, pp. 14–19, 2015.

[14] A. Bhattacharya, W. Wu, and Z. Yang, "Quality of experience evaluation of voice communication: an affect-based approach," *Human-centric Computing and Information Sciences*, vol. 2, no. 1, 2012.

[15] S. Porcu, A. Floris, and L. Atzori, "Towards the Prediction of the Quality of Experience from Facial Expression and Gaze Direction," in *2019 22nd Conf. on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*, Feb 2019, pp. 82–87.

[16] S. Porcu, S. Uhrig, J. Voigt-Antons, S. Möller, and L. Atzori, "Emotional Impact of Video Quality: Self-Assessment and Facial Expression Recognition," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, June 2019, pp. 1–6.

[17] J.-N. Antons, R. Schleicher, S. Arndt, S. Möller, A. K. Porbadnigk, and G. Curio, "Analyzing speech quality perception using electroencephalography," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 721–731, 2012.

[18] S. Uhrig, G. Mittag, S. Möller, and J.-N. Voigt-Antons, "P300 indicates context-dependent change in speech quality beyond phonological change," *Journal of Neural Engineering*, 2019.

[19] S. Arndt, J.-N. Antons, R. Schleicher, and S. M "Using electroencephalography to analyze sleepiness due to low-quality audiovisual stimuli," *Signal Processing: Image Comm.*, vol. 42, pp. 120 – 129, 2016.

[20] E. Kroupi, P. Hanhart, J. Lee, M. Rerabek, and T. Ebrahimi, "EEG correlates during video quality perception," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 2135–2139.

[21] S. Arndt, J. Radun, J.-N. Antons, and S. Möller, "Using eye-tracking and correlates of brain activity to predict quality scores," in *Proc. of the Sixth Int. Workshop on Quality of Multimedia Experience (QoMEX), 2014*. IEEE, 2014, pp. 281–285.

[22] Y. Rai, M. Barkowsky, and P. L. Callet, "Role of spatio-temporal distortions in the visual periphery in disrupting natural attention deployment," *Electronic Imaging*, vol. 2016, no. 16, pp. 1–6, 2016.

[23] P. Rai, Yashasand Le Callet, "Do gaze disruptions indicate the perceived quality of non-uniformly coded natural scenes? ," in *Human Vision and Electronic Imaging 2017*, 2017, pp. 104–109.

[24] U. Engelke, R. Pepion, P. L. Callet, and H.-J. Zepernick, "Linking distortion perception and visual saliency in H.264/AVC coded video containing packet loss," in *Proc. SPIE*, vol. 7744, 2010.

[25] J. N. Bailenson, E. D. Pontikakis, I. B. Mauss, J. J. Gross, M. E. Jabon, C. A. Hutcherson, C. Nass, and O. John, "Real-time classification of evoked emotions using facial feature tracking and physiological responses," *Int. Journal of Human-Computer Studies*, vol. 66, no. 5, pp. 303 – 317, 2008.

[26] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal Trajectory Aware Video Quality Measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 266–279, 2009.

[27] S. Mustafa and A. Hameed, "Perceptual quality assessment of video using machine learning algorithm," *Signal, Image and Video Processing*, vol. 13, no. 8, pp. 1495–1502, 2019.

[28] M. Torres Vega, D. C. Mocanu, S. Stavrou, and A. Liotta, "Predictive no-reference assessment of video quality," *Signal Processing: Image Communication*, vol. 52, pp. 20–32, 2017.

[29] M. Seufert, P. Casas, N. Wehner, L. Gang, and K. Li, "Features that Matter: Feature Selection for On-line Stalling Prediction in Encrypted Video Streaming," in *IEEE INFOCOM 2019 - IEEE Conf. on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019, pp. 688–695.

[30] D. C. Mocanu, H. Bou Ammar, D. Lowet, K. Driessens, A. Liotta, G. Weiss, and K. Tuyls, "Factored four way conditional restricted Boltzmann machines for activity recognition," *Pattern Recognition Letters*, vol. 66, pp. 100 – 108, 2015, pattern Recognition in Human Computer Interaction.

[31] D. Minovski, C. Åhlund, K. Mitra, and P. Johansson, "Analysis and Estimation of Video QoE in Wireless Cellular Networks using Machine Learning," in *2019 Eleventh Int. Conf. on Quality of Multimedia Experience (QoMEX)*, 2019, pp. 1–6.

[32] P. Casas, A. D'Alconzo, F. Wamser, M. Seufert, B. Gardlo, A. Schwind, P. Tran-Gia, and R. Schatz, "Predicting QoE in cellular networks using machine learning and in-smartphone measurements," in *2017 Ninth Int. Conf. on Quality of Multimedia Experience (QoMEX)*, 2017, pp. 1–6.

[33] P. Charonyktakis, M. Plakia, I. Tsamardinos, and M. Papadopouli, "On user-centric modular QoE prediction for VoIP based on machine-learning algorithms," *IEEE Trans. on mobile computing*, vol. 15, no. 6, pp. 1443–1456, 2015.

[34] E. Cipressi and M. L. Merani, "An Effective Machine Learning (ML) Approach to Quality Assessment of Voice over IP (VoIP) Calls," *IEEE Networking Letters*, pp. 1–1, 2020.

[35] M. Lopez-Martin, B. Carro, J. Lloret, S. Egea, and A. Sanchez-Esguevillas, "Deep Learning Model for Multimedia Quality of Experience Prediction Based on Network Flow Packets," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 110–117, 2018.

[36] A. Lekharu, K. Y. Moulii, A. Sur, and A. Sarkar, "Deep Learning based Prediction Model for Adaptive Video Streaming," in *2020 International Conference on COMmunication Systems NETworkS (COMSNETS)*, 2020, pp. 152–159.

[37] M. Torres Vega, D. C. Mocanu, J. Famaey, S. Stavrou, and A. Liotta, "Deep Learning for Quality Assessment in Live Video Streaming," *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 736–740, 2017.

[38] N. Barman, E. Jammeh, S. A. Ghorashi, and M. G. Martini, "No-Reference Video Quality Estimation Based on Machine Learning for Passive Gaming Video Streaming Applications," *IEEE Access*, vol. 7, pp. 74 511–74 527, 2019.

[39] M. Torres Vega, C. Perra, F. De Turck, and A. Liotta, "A Review of Predictive Quality of Experience Management in Video Streaming Services," *IEEE Trans. on Broadcasting*, vol. 64, no. 2, pp. 432–445, June 2018.

[40] A. T. Lopes, E. de Aguiar, A. F. D. Souza, and T. Oliveira-Santos, "Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610 – 628, 2017.

[41] S. Xie and H. Hu, "Facial Expression Recognition Using Hierarchical Features With Deep Comprehensive Multipatches Aggregation Convolutional Neural Networks," *IEEE Trans. on Multimedia*, vol. 21, no. 1, pp. 211–220, Jan 2019.

[42] H. Kim, Y. Kim, S. J. Kim, and I. Lee, "Building Emotional Machines: Recognizing Image Emotions Through Deep Neural Networks," *IEEE Trans. on Multimedia*, vol. 20, no. 11, pp. 2980–2992, Nov 2018.

[43] C. Ding and D. Tao, "Robust Face Recognition via Multimodal Deep Face Representation," *IEEE Trans. on Multimedia*, vol. 17, no. 11, pp. 2049–2058, Nov 2015.

[44] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A Deep Neural Network-Driven Feature Learning Method for Multi-view Facial Expression Recognition," *IEEE Trans. on Multimedia*, vol. 18, no. 12, pp. 2528–2536, Dec 2016.

[45] R. Schleicher and J.-N. Antons, "Evoking Emotions and Evaluating Emotional Impact," in *Quality of Experience: Advanced Concepts, Applications and Methods*, S. Möller and A. Raake, Eds. Cham: Springer International Publishing, 2014, pp. 121–132.

[46] D. Ghadiyaram, A. C. Bovik, H. Yeganeh, R. Kordasiewicz, and M. Gallant, "Study of the effects of stalling events on the quality of experience of mobile streaming videos," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 989–993.

[47] ——. (2016) LIVE Mobile Stall Video database. [Online]. Available: http://live.ece.utexas.edu/research/LIVEStallStudy/index.html

[48] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via Crowdsourcing," in *2011 IEEE Int. Symposium on Multimedia*, 2011, pp. 494–499.

[49] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best Practices for QoE Crowdtesting: QoE Assessment With Crowdsourcing," *IEEE Trans. on Multimedia*, vol. 16, no. 2, pp. 541–558, 2014.

[50] M. Hirth, T. Hoßfeld, M. Mellia, C. Schwartz, and F. Lehrieder, "Crowdsourced network measurements: Benefits and best practices," *Computer Networks*, vol. 90, pp. 85 – 98, 2015.

[51] A. Raake, M.-N. Garcia, W. Robitza, P. List, S. Göring, and B. Feiten, "A bitstream-based, scalable video-quality model for HTTP adaptive streaming: ITU-T P.1203.1," in *Ninth Int. Conf. on Quality of Multimedia Experience (QoMEX)*. IEEE, May 2017.

[52] T. Hoßfeld, P. E. Heegaard, M. Varela, and L. Skorin-Kapov, "Confidence Interval Estimators for MOS Values," 2018.

[53] S. M. Safdarnejad, X. Liu, L. Udpa, B. Andrus, J. Wood, and D. Craven, "Sports Videos in the Wild (SVW): A video dataset for sports analysis," in *2015 11th IEEE Int. Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1, 2015, pp. 1–7.

[54] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *2018 13th IEEE Int. Conf. on Automatic Face Gesture Recognition (FG 2018)*, 2018, pp. 59–66.

[55] E. Wood, T. Baltrušaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling, "Rendering of Eyes for Eye-Shape Registration and Gaze Estimation," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3756–3764.

[56] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic Action Unit detection," in *2015 11th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 06, 2015, pp. 1–6.

[57] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. San Francisco: Consulting Psychologists Press, 1978.

[58] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, June 2010, pp. 94–101.

[59] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, June 2008, pp. 1322–1328.

[60] Haibo He, Yang Bai, E. A. Garcia, and Shutao Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322–1328.

[61] F. Crété-Roffet, T. Dolmiere, P. Ladret, and M. Nicolas, "The Blur Effect: Perception and Estimation with a New No-Reference Perceptual Blur Metric," in *SPIE Electronic Imaging Symposium Conf Human Vision and Electronic Imaging*, 2007, pp. EI 6492–16.

[62] D. Q. Bao. (2020) Image Blur Metric, MATLAB Central File Exchange.

[63] ITU, "Models and tools for quality assessment of streamed media." Recommendation ITU-T P.1203, 2017.

[64] W. Robitza, S. Göring, A. Raake, D. Lindegren, G. Heikkilä, J. Gustafsson, P. List, B. Feiten, U. Wüstenhagen, M.-N. Garcia, K. Yamagishi, and S. Broom, "HTTP Adaptive Streaming QoE Estimation with ITU-T Rec. P.1203 – Open Databases and Software," in *9th ACM Multimedia Systems Conference*, 2018.



Simone Porcu is a PhD student at the department of Electrical and Electronic Engineering at the University of Cagliari (Italy) from the 2017. S. Porcu received a bachelor's degree in Computer science at the University of Cagliari in 2015 and a Master degree in Computer Science in 2017. He is a vice-chair of IEEE Cagliari Student Branch. S. Porcu research interests are in Multimedia Communication focused on multimedia Quality of Experience, Machine Learning, Image Analysis and Human Behaviour.



Alessandro Floris is an Assistant Professor at the Department of Electrical and Electronic Engineering (DIEE), University of Cagliari (Italy). He received his M. Sc. degree in Electronic Engineering and his Ph. D. in Electronic and Computer Engineering from University of Cagliari respectively in 2011 and 2017. He is a member of the Multimedia & Communications Laboratory (MCLab) (mclab.diee.unica.it.) and of the Italian University Consortium for Telecommunications (CNIT). His main research interests include Quality of Experience (QoE) evaluation and management for multimedia services, Internet of Things (IoT) systems and Smart City applications.



Jan-Niklas Voigt-Antons received his diploma in psychology in 2008 from the Technische Universität Darmstadt, Germany, a Doctor-of-Engineering degree in 2014 from the Technische Universität Berlin, Germany and has been researching at the Quality and Usability Lab of Technische Universität (TU) Berlin, since. His research interests are in Quality-of-Experience evaluation, and it's physiological correlates with an emphasis on media transmissions and human-machine-interaction, including neural processing of multimodal interaction. During summer 2012, he visited the researcher at MuSAE Lab (INRS-EMT), Canada, where he examined the neural correlations of quality perception for complex speech signals. In spring 2014, he was visiting researcher at the department of psychology of NTNU, Norway where he examined neural correlates of audiovisual asynchrony.



Luigi Atzori is Full Professor at the Department of Electrical and Electronic Engineering at the University of Cagliari (Italy) and Research Associate at the Multimedia Communications Laboratory of CNIT (Consorzio Nazionale Interuniversitario per le Telecomunicazioni). L. Atzori research interests are in multimedia communications and computer networking (wireless and wireline), with emphasis on multimedia QoE, multimedia streaming, NGN service management, service management in wireless sensor networks, architecture and services in the Internet of Things. L. Atzori has been the coordinator of the Marie Curie Initial Training Network on QoE for multimedia services (qoenet-itn.eu, 2015-2018), involving ten European Institutions in Europe and one in South Korea. He has been the editor for the ACM/Springer Wireless Networks Journal and guest editor for the IEEE Communications Magazine, the Springer Monet Journal and the Elsevier Signal Processing: Image Communications Journal. He is member of the editorial board of the IEEE IoT, the Elsevier Ad Hoc Networks and the Advances on Multimedia journals.



Sebastian Möller studied Electrical Engineering at the universities of Bochum (Germany), Orléans (France) and Bologna (Italy). He received a Doctor-of-Engineering degree in 1999 with a book on the assessment and prediction of speech quality in telecommunications, and the Venia Legendi with a book on the quality of telephone-based spoken dialogue systems in 2004. In 2005, he joined Telekom Innovation Laboratories, and in 2007, he was appointed full professor for Quality and Usability at TU Berlin. In addition, he is Scientific Director at the German Research Center for Artificial Intelligence (DFKI) in Berlin, and Adjunct Professor at the University of Technology Sydney. His primary interests are in speech and multimedia quality, speech and language technology, user experience, and crowdsourcing. Since 1997, he has taken part in ITU–T Study Group 12, where he is currently Co-Rapporteur of Question Q.15/12.