Università degli Studi di Cagliari

PHD DEGREE

Electronic and Computer Engineering

Cycle XXXIII

TITLE OF THE PHD THESIS

Estimation of the QoE for video streaming services based on facial expressions and gaze direction

Scientific Disciplinary Sector(s)

S.S.D.ING-INF/03

PhD Student: Simone Porcu

Supervisor Prof. Luigi Atzori, PhD Alessandro Floris

Final exam. Academic Year 2019 – 2020
Thesis defence: February 2021 Session

University of Cagliari

Department of Electrical and Electronic Engineering
PhD Course in Electronic and Computer Engineering
Cycle XXXIII

Ph.D. Thesis

Estimation of the QoE for video streaming services based on facial expressions and gaze direction

S.S.D. ING-INF/03

Candidate
Simone Porcu

PhD Supervisor
Prof. Luigi Atzori
PhD Alessandro Floris

PhD Coordinator
Prof. Alessandro Giua

Final examination academic year 2019/2020

"There's no such thing as a free lunch"

I want to state a particular thank to prof. Luigi Atzori and Alessandro, who have always supported me in the realization of this work.

Contents

Acronyms

ACR Absolute Category Rating

AI Artificial Intelligence

AU Action Units

CNN Convolutional Neural Network

CR cropping

DA Data Augmentation

DL Deep Learning

ECG Electrocardiogram

EEG Electroencephalogram

FACS Facial Action Coding System

FER Facial Emotion Recognition

GAN Generative Adversarial Network

HR horizontal reflection

IL impairment level

K-NN K-Nearest Neighbors

KQIs Key Quality Indicators

LR Linear Regression

ML Machine Learning

MLR Multiple Linear Regression

QoE Quality of Experience

QoS Quality of Service

ReLU Rectified Linear Unit

ROI regions of interest

SVM Support Vector Machine

TR translation

DT Decision Tree

VR vertical reflection

Abstract

As the multimedia technologies evolve, the need to control their quality becomes even more important making the Quality of Experience (QoE) measurements a key priority. Machine Learning (ML) can support this task providing models to analyse the information extracted by the multimedia. It is possible to divide the ML models applications in the following categories:

- *QoE modelling*: ML is used to define QoE models which provide an output (e.g., perceived QoE score) for any given input (e.g., QoE influence factor).
- *QoE monitoring in case of encrypted traffic*: ML is used to analyze passive traffic monitored data to obtain insight into degradations perceived by end users.
- *Big data analytics*: ML is used for the extraction of meaningful and useful information from the collected data, which can further be converted to actionable knowledge and utilized in managing QoE.

The QoE estimation quality task can be carried out by using two approaches: the objective approach and subjective one. As the two names highlight, they are referred to the pieces of information that the model analyses. The objective approach analyses the objective features extracted by the network connection and by the used media. As objective parameters, the state-of-the-art shows different approaches that use also the features extracted by human behaviour. The subjective approach instead, comes as a result of the rating approach, where the participants were asked to rate the perceived quality using different scales. This approach had the problem to being a time-consuming approach and for this reason not all the users agree to compile the questionnaire. Thus the direct evolution of this approach is the ML model adoption. A model can substitute the questionnaire and evaluate the QoE, depending on the data that analyses. By modelling the human response to the perceived quality on multimedia, QoE researchers found that the parameters extracted from the users could be different, like Electroencephalogram (EEG), Electrocardiogram (ECG), waves of the brain. The main problem with these techniques is the hardware. In fact, the user must wear electrodes in case of ECG and EEG, and also if the obtained results from these methods are relevant, their usage in a real context could be not feasible. For this reason my studies have been focused on the

developing of a Machine Learning framework completely unobtrusively based on the Facial reactions.

Introduction

The 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, required a deeper control for the network multimedia transmissions. Indeed, Cisco predicts global IP video traffic to be 82% of all IP traffic by 2022, up from 75% in 2017 [Cis20]. 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 [SKVHC18, AA20]. Accordingly, the 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 [LKB$^+$19]. QoE is defined as the 'the degree of delight or annoyance of the user of an application or service.' [LCMP12]. 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 [ITU08].

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 [CDF$^+$14, KDSS19]. Indeed, besides coding and network distortions that are measurable, there are other factors that may influence the user's experience, such as the user's characteristics, the device and the context of use, which may vary depending on the user [CSW$^+$16, YLG$^+$18]. However, while convenient and effective, this kind of self-report technique can be problematic because it may be subject to bias from factors not related to the stimulus. These may include, for example, the interviewer's reaction to the questions, the way the questions are formulated, and the context (tests are typically conducted in the laboratory). Additionally, users may be annoyed by surveys and interviews, which are generally boring and time-consuming. For these reasons, alternative approaches for QoE evaluation started to be investigated, such as psychophysiological measures (EEG, gaze direction) and facial expressions [EDM$^+$17, ABSM18]. The subject of this research's thesis 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. The QoE evaluation approach just introduced is the outcome of three years of studies, experiments and research on the QoE and Facial Emotion Recognition (FER) field. FER field or more general the expression recognition field are a not well explored field. The facial emotion detection has been approached in different ways. The state of the art shows two main approach to understand the facial expressions. The analysis could be based on Facial Action Coding System (FACS), that analyses the movement of each muscle of the face, or the facial expression usually analysed with a Convolutional Neural Network (CNN). Both approaches use two branches of the Artificial Intelligence (AI) as key to understanding the facial reactions. The usage of a ML model or a CNN permits to reach results that in another way could be impossible to obtain. CNNs permit to make deeper analysis on images to extract the characteristics, but they require a huge amount of data to be perfectly trained. ML models can be used as well as the CNNs but from our studies, they cannot reach comparable results to CNNs. Their strength is the lightness and efficiency in smaller problems obtaining high performance in accuracy and timing terms. The AI also cover different research application areas because of its dynamism. In fact, just talking about the QoE field, ML models helped researchers to correlate the Quality of Service (QoS) key factor to the level of pleasure of the final user [ABSM18]. Moreover, as it will be explained in chapter 2, a lot of experiments have been investigated analysing the human physiology. All the humans' extracted pieces of information have been correlated using statistical methods and also using ML models and CNNs. Their strength relies on the capability in making pattern recognition between the information, making easier the work for the researcher. A deeper explanation about these two branches of the AI is given in part 5. Anyway, both approaches are applied 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 [TDL+15, ABSM18, AWZ12]. In my research both methods have been deeply studied, obtaining the best results whit the FACS features approach, but for clarity all the workflow regarding the CNN approach and the FACS method is explained in the following chapters.

The dissertation is organized as follows:

- Part 1 shows the different developed techniques used to increase the recognition accuracy of the facial expressions using Convolutional Neural Networks and a Generative Adversarial Network.

- Part 2 explains the AI framework for the video-streaming quality estimation based on facial expressions. Based on the previous studies about the FER, in this part the transition from the FER general approach to the more precisely FACS is explained.

- Part 3 shows the application of the introduced approach to a smart-room scenario, giving also a definition of a Quality of Experience Managements system for Smart City services.

- Part 4 shows the final conclusions and future works.

Part I

Facial Expression

Chapter 1

Facial Expression Recognition

1.1 Introduction

As explained in the previous chapter, an approach to understanding the perceived quality from the facial expressions is the final goal of the whole research. The first study has been focused on the application of a CNN to understand the facial expressions. Its result has been correlated with the perceived quality. To understand its feasibility I started to investigate what the facial expressions were. The State of the Art explains that Facial Expression Recognition (FER) is a challenging task involving scientists from different research fields, such as psychology, physiology and computer science, whose importance has been growing in the last years due to vast areas of possible applications, e.g., human-computer interaction, gaming, healthcare. The six basic emotions (i.e., anger, fear, disgust, happiness, surprise, sadness) were identified by Ekman and Freisen as main emotional expressions that are common among human beings [EF71]. The CNN usage is nowadays a common approach to face the FER problem. The CNNs thanks to their structure are efficient methods to analyse or make pattern recognition on the images. The main problem that the CNN usage highlights, it is the huge amount of data that it requires to be trained in the proper way. Therefore, researches adopted a lot of Data Augmentation (DA) techniques to increase the number of images. In this part, we will focus on the analysis of these DA techniques. As the main objective of this research is to investigate whether the selection of the proper DA technique may compensate the lack of large DB for training the FER model, this work, not provides a novel CNN but the whole work has been based on the VGG16 CNN.

The main reasons at the basis of the proposed study are: i) DA techniques for FER systems are mostly used to remedy the small dimension of public image-labelled DBs. However, it is not always easy to understand which DA technique may be more convenient for FER systems because most of the state-of-the-art experiments use different settings which makes the impact of DA techniques not comparable. ii) There is limited research in this regard. To the best of the author knowledge, the only study proposing a comparison of DA techniques is [PWBL17]. However, only limited DA techniques are considered, smaller DBs are used, and no cross-database evaluation is performed. iii) The specific utilization of GAN as a DA technique for FER systems is barely investigated in the literature.

1.2 Background

Most of traditional FER studies consider the combination of face appearance descriptors, which are used to represent facial expressions, with deep learning techniques to handle the challenging factors for FER, achieving the state-of-the-art recognition accuracy [LD18]. The study in [GAC16] used Principal Component Analysis (PCA), Local Binary Pattern (LBP) histogram and Histogram of Oriented Gradient (HOG) for sample image representation whereas the proposed FER system is based on boosted neural network ensemble (NNE) collections. Experimental results are accomplished on multicultural facial expression datasets (RaFD, JAFFE, and TFEID) and the achieved accuracy is comparable to state-of-the-art works. A novel edge-based descriptor, named Local Prominent Directional Pattern (LPDP), is proposed in [MAAWI$^+$19]. The LPDP considers statistical information of a pixel neighborhood to encode more meaningful and reliable information than the existing descriptors for feature extraction. Extensive experiments on FER on well-known datasets (CK+, MMI, BU-3DFE, ISED, GEMEP-FERA, FACES) demonstrate the better capability of LPDP than other existing descriptors in terms of robustness in extracting various local structures originated by facial expression changes. Shan et al. [SGM09] used LBP as feature extractor and combined different machine learning techniques to recognize facial expressions. The best result are obtained by using LBP and Support Vector Machine (SVM) on the CK+ DB. Cross-database validation is also performed training with the CK+ and testing with JAFFE. In [LMR$^+$17], a FER system is proposed based on an automatic and more efficient facial decomposition into regions of interest (ROI). These ROIs represent 7 facial components involved in the expression of emotions (left eyebrow, right eyebrow, left eye, right eye, between eyebrows, nose and mouth) and are extracted using the positions of some landmarks. A multiclass SVM classifier is then used to classify the six basic facial expressions and the neutral state. A cross-database evaluation is also performed training with the KDEF DB and testing with the CK+ DB. Similarly, Gu *et al.* [GXV$^+$12] divided each image into several local ROIs containing facial components critical for recognizing expressions (i.e., eyebrow corner, mouth corner and wrinkle). Each of these ROIs is then subjected to a set of Gabor filters and local classifiers that produce global features representing facial expressions. In-group and cross-database experiments are conducted using CK+ and JAFFE DBs. In [AD14], landmark points are used for the recognition of the six basic facial expressions in images. The proposed technique relies on the observation that the vectors formed by the landmark point coordinates belong to a different manifold for each of the expressions. Extensive experiments have been performed on two publicly available datasets (MUG and CK+) yielding very satisfactory expression recognition accuracy. [dSP15] investigates the performance of cross-classification using facial expression images from different cultures taken from 4 DBs, i.e., CK+, MUG, BOSPHOROUS and JAFFE. Different combinations of descriptors (HOG, Gabor filters, LBPs) and classifier (SVM, NNs, k-NNs) were employed for experiments. Experiment results highlighted that the most effective combination for in-group classification was formed by the association of HOG filter and SVM. However, when classifying different DBs, even with the most effective combination, the accuracy dropped considerably. [HM17] proposed a novel Deep Neural Network (DNN) method for FER based on a two-step learning process aimed to predict the labels of each frame while considering the labels of adjacent frames in the sequence. Experimental evaluations performed with three DBs (CK+, MMI, and FERA) showed that the proposed network outperforms the state-of-the-art methods in cross-database tasks. [MCM16] presented a novel deep neural network architecture for the FER problem, which examines the network's ability to perform cross-database classification. Experiments are conducted on seven well-known facial expressions DBs (i.e., MultiPIE, MMI, CK+, DISFA, FERA, SFEW, and FER2013) obtaining results better than, or comparable to, state-of-the-art methods. Lopes *et al.* [LdASOS17] propose a combination of CNN and image pre-processing steps aimed to reduce the need for a large amount of data by decreasing the variations between images selecting a subset of the features to be learned. The experiments were carried out using three public DBs (i.e., CK+, JAFFE and BU-3DFE). Both in-group and cross-database evaluations are performed.

Although numerous studies have been conducted on FER, it remains one of the hardest tasks for image classification systems due to the following main reasons: i) significant overlap between basic emotion classes [EF71]; ii) differences in the cultural manifestation of emotion [dSP15]; iii) need of a large amount of training data to avoid overfitting [LD18].

Moreover, many state-of-the-art FER methods present a misleading high-accuracy because no cross-database evaluations are performed. Indeed, facial features from one subject in two different expressions can be very close in the features space; conversely, facial features from two subjects with the same expression may be very far from each other [LdASOS17]. For these reasons, cross-database analyses are preferred to improve the validity of FER systems, i.e., training the system with one DB and testing with another one.

Large DBs are typically needed for training and testing machine learning algorithms, and in particular deep learning algorithms intended for image classification systems, mostly to avoid overfitting. However, although there are some public image-labelled DBs widely used to train and test FER systems, such as the Karolinska Directed Emotional Faces (KDEF) [GDRLV08] and the Extended Cohn-Kanade (CK+) [LCK^{+}10], these may not be large enough to avoid overfitting. Techniques such as DA are then commonly used to remedy the small dimension and/or class imbalance of these public DBs by increasing the number of training samples. Basically, DA techniques can be grouped into two main types [SK19]: i) data warping augmentations: generate image data through label-preserving linear transformations, such as geometric (e.g., translation, rotation, scaling) and colour transformations [SSP03, LHZL18]; ii) oversampling augmentations: task-specific or guided-augmentation methods which create synthetic instances given specific labels (e.g., Generative Adversarial Network (GAN)) [YSH18, ZLL^{+}18]. Although the second can be more effective, their computational complexity is higher.

In literature are presented few works that show the efficiency of this approach. Due to the small dimension of public image-labelled DBs, DA techniques are commonly used to augment the DB dimension. Geometric DA techniques are used the most, due to their low computational complexity. Simard *et al.* demonstrated the benefits of applying geometric transformations of training images for DA, such as translations, rotations and skewing [SSP03]. In [LHZL18], five DA geometric techniques have been included in the FER system, which are rotation, shearing, zooming, horizontal flip, and rescale. Experimental results showed that DA boosted the model in terms of accuracy. In [ZBO17], two DA techniques are considered in the proposed FER system, i.e., random noise and skew. The former adds random noise to the position of the eyes whereas the latter applies a random skew, i.e., changes the corners of the image to generate distortion. Extensive cross-database experimentations are conducted using 6 DBs (i.e., CK+, JAFFE, MMI, RaFD, KDEF, BU3DFE, ARFace), which outperform state-of-the-art results in most of the cases. In [PWBL17], four DA methods are compared as CNN enhancer: resize, face detection & cropping, adding noises, and data normalization. The combination of face detection with adding noises for DA boosted the performance of the CNN in terms of accuracy.

In [PWBL17], four DA methods are compared as CNN enhancer: resize, face detection & cropping, adding noises, and data normalization. The combination of face detection with adding noises for DA boosted the performance of the CNN in terms of accuracy. Besides geometric transformations, more complex guided-augmentation methods may be used for DA, such as the GAN [GPAM^{+}14]. In [ZSX^{+}19], a general framework of DA using GANs in feature space for imbalanced classification is proposed. Experiments were conducted on three DBs, i.e., SVHN, FER2013, and Amazon Review of Instant Video, which showed significant improvement with feature augmentation of GANs. In [YSH18], a conditional GAN is used to generate images aimed at augmenting the FER2013 dataset. A CNN is used for training the prediction model and the average accuracy obtained 5% increase after adopting the GAN DA technique. In [ZLL^{+}18], a framework is designed using a CNN model as the classifier and a cycle-consistent adversarial networks (CycleGAN) as the generator for DA. Experiments on 3 DBs (FER2013, JAFFE, SFEW) showed that the proposed GAN-based DA technique can obtain $5\% - 10\%$ increase in the accuracy. In [CHW^{+}19], a FER method based on Contextual GAN is proposed, which uses a contextual loss function to enhance the facial expression image and a reconstruction

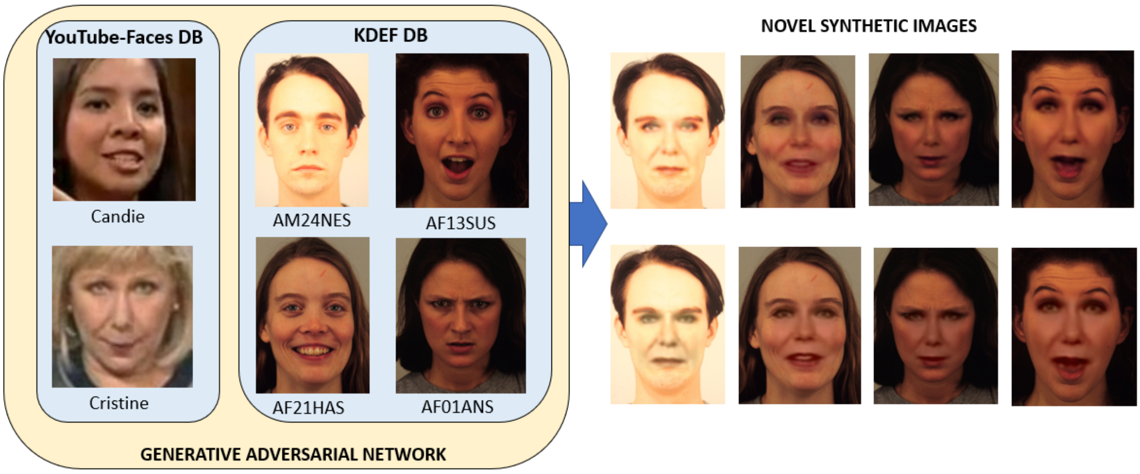


Figure 1.1: Examples of novel synthetic images generated with the GAN.

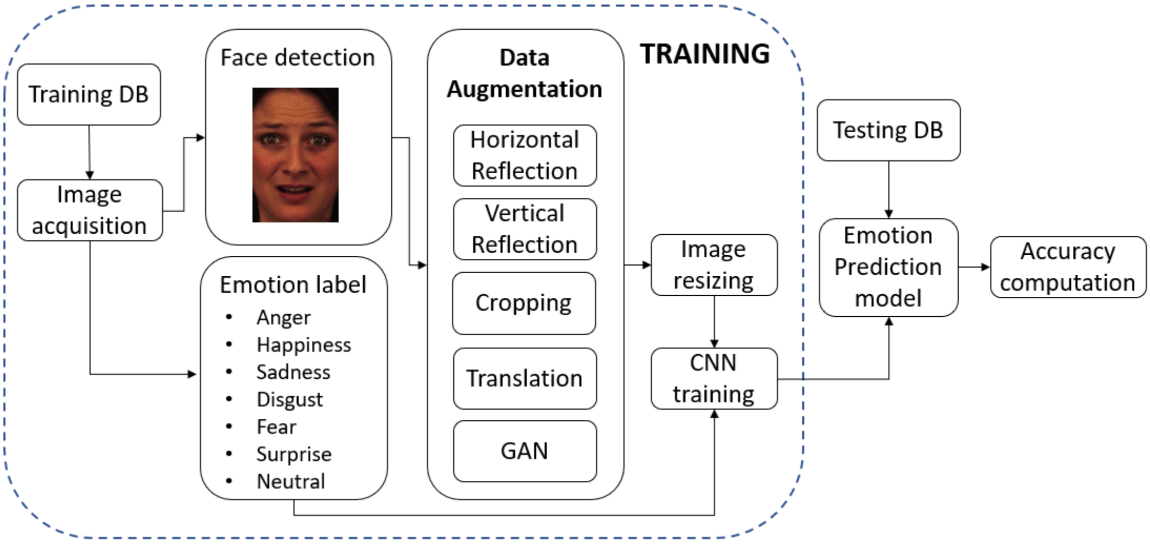


Figure 1.2: The proposed FER system. The dashed line separates the training phase from the testing phase.

loss function to maintain the identity information of the subject in the expression image. Experimental results on the augmented CK+ and KDEF DBs show that this method improves the recognition accuracy of about 5%. However, all these studies have not performed a cross-database evaluation.

1.3 The proposed FER system

The framework of the proposed FER system is shown in Fig. 1.2. The dashed line separates the training phase from the testing phase. The operations performed during the training phase are the following. Firstly, a face detection operation is performed to detect and select only the face information from the training images, which are acquired from the selected training DB. The emotion labels associated to each training image are used by the CNN to train the Emotion Prediction model. Different DA techniques are implemented to augment the training DB, namely: horizontal reflection (HR),vertical reflection (VR), cropping (CR), translation (TR), GAN. These techniques have the objective to increase the size of the training DB for the model training process by producing novel but different versions of the original images. The augmented images are resized to be compatible with the CNN accepted image size. The CNN is trained with these images and its output is the Emotion Prediction model, which is able to process a face image and predict one of the six basic human emotions (anger, sadness, surprise, happiness, disgust and fear) or the neutral emotion. The experiment design aims to identify the impact of each DA technique on the accuracy of the Emotion Prediction model. The performance of the Emotion Prediction model is computed using the testing DB, which does not contain any image used for training the model so as to perform cross-DB evaluation.

In the following sections, details of the major elements of the proposed FER system will be provided.

1.3.1 Face detection

The face detection operation allows selecting only the face information from the input image by removing all unnecessary data for emotion recognition. Firstly, the RGB input image is converted into a gray-scale image. Then, for the face detection operation the S^{3}FD has been used: Single Shot Scale-invariant Face Detector [ZZL$^+$17], which solves the problem of anchor-based detection methods whose performance decrease rapidly as the faces becoming smaller.

1.3.2 Geometric DA techniques

- *Horizontal reflection*: the HR, also known as horizontal flip, is a DA technique that creates a mirrored image from the original one along the vertical direction.

- *Vertical reflection*: the VR, also known as vertical flip, is a DA technique that creates a mirrored image from the original one along the horizontal direction. A VR is equivalent to rotating an image by 180 degrees and performing a HR.

- *Translation*: the TR DA technique performs a random moving of the original image along the horizontal or vertical direction (or both). Padding zeros are added to the image sides.

- *Cropping*: the CR DA technique randomly samples a section from the original image. The cropped image size is large enough to contain a relevant part of the face.

1.3.3 Generative Adversarial Network

A GAN is a framework including two deep networks, one (the generative) pitted against the other (the discriminative). This approach allows the neural network to create new data with the same distribution of training data. The generator attempts to produce a realistic image to fool the discriminator, which tries to distinguish whether this image is from the training set or the generated set [ZLL$^+$18]. In the proposed FER system, I used the GAN implemented for the DeepFake autoencoder architecture of the FaceSwap project[1]. In Fig. 1.1, I show some novel synthetic images generated with the GAN. The face images from the KDEF DB are used as the base to create novel synthetic images using the facial features of 2 images (i.e., Candie Kung and Cristina Saralegui) selected from the YouTube-Faces DB [WHM11]. It can be seen how the novel images differ between each other, in particular with respect to the eyes, the nose, and the mouth, whose characteristics are taken from the Candie and Cristine images. A DB with these novel synthetic images has been created and made it public[2].

1.3.4 Convolutional Neural Network

The CNN considered for the proposed FER model is the VGG16, a popular CNN proposed by K. Simonyan and A. Zisserman, which competed in the ImageNet Large Scale Visual Recognition Challenge achieving a top-5 accuracy of 92.7% [SZ14]. The VGG16 has made the improvement over the AlexNet CNN by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 33 kernel-sized filters one after another.

1.4 Experimental results

The experiments were performed using 3 publicly available DBs in the FER research field: the KDEF [GDRLV08], the CK+ [LCK$^+$10] and the ExpW [ZLLT16]. The KDEF DB is a set of 4900 pictures of human facial expressions with associated emotion label. It is composed of 70 individuals, each displaying the 6 basic emotions plus the neutral emotion. Each emotion was photographed (twice) from five different angles, but for our experiments I only used the frontal poses, for a total of 490 pictures. The CK+ DB includes 593 video sequences recorded from 123 subjects. Among these videos, 327 sequences from 118 subjects are labeled with one of the

[1] https://github.com/deepfakes/faceswap

[2] http://mclab.diee.unica.it/?p=272

Data Augmentation	**Testing DB: CK+**	**Testing DB: ExpW**
	Accuracy	**Accuracy**
No DA	53 %	15.7 %
CR	50 %	-
VR	53 %	-
HR	57 %	-
TR	61 %	-
GAN	75 %	-
HR & TR	63 %	20 %
GAN & HR & TR	**83.2** %	**43.2** %

Table 1.1: Comparison of the overall emotion recognition accuracy achieved by the proposed FER system considering different DA techniques. Training DB: KDEF, Testing DBs: CK+ and ExpW.

6 basic emotions plus the neutral emotion. The ExpW DB contains 91,793 faces labeled with the 6 basic emotions plus the neutral emotion. This DB contains a quantity of images much larger than the others DB and with more diverse face variations.

A cross-database evaluation process was performed by training the proposed FER model with the KDEF DB augmented with the considered DA techniques with the following setting: 1) 490 images created with HR; 2) 490 images created with VR; 3) 490 images created randomly translating 20% of each image (optimal setting found empirically); 4) 490 images created randomly cropping 50% of each image (optimal setting found empirically); 5) 980 images generated with the GAN (70 individuals from KDEF DB × 7 emotions × 2 subjects from YouTube-Faces DB). The size of the training DB was 980 images when using HR, VR, TR and CR DA techniques, and 1470 images when using the GAN.

The DA techniques as well as the CNN have been implemented with PyTorch [PGC$^+$17] and the related libraries. The experiments were conducted with a Microsoft Windows 10 machine with the NVIDIA CUDA Framework 9.0 and the cuDNN library installed. All the experiments were carried out using an Intel Core i9-7980XE with 64 GB of RAM memory and two graphic cards Nvidia GeForce GTX 1080 Ti with 11 GB of GPU memory each one.

The testing has been performed with the CK+ (first experiment) and ExpW DBs (second experiment). Table 1.1 summarizes the values of the overall emotion recognition accuracy achieved considering different DA techniques. Specifically, I computed the macro-accuracy; however, I refer to it in the rest of the work only as accuracy. In Table 1.2, I compare the recognition accuracy achieved for the single emotions for the best combination of DA techniques.

Data Augmentation: GAN & HR & TR	**Testing DB: CK+**	**Testing DB: ExpW**
	Accuracy	**Accuracy**
Anger	50 %	20.7 %
Sadness	100 %	28 %
Surprise	60 %	56.2 %
Happiness	100 %	62.5 %
Disgust	100 %	20.3 %
Fear	75 %	20.3 %
Neutral	100 %	88.2 %

Table 1.2: Comparison of the recognition accuracy of single emotions achieved by the proposed FER system for the best combination of DA techniques, i.e., GAN & HR & TR. Training DB: KDEF, Testing DBs: CK+ and ExpW.

1.4.1 First experiment: Testing with the CK+ DB

In the first experiment (testing with the CK+ DB), when no DA techniques are used, i.e., the FER system is trained with the original dataset, the achieved overall accuracy is 53%. The CR is the only DA technique that causes a reduction of the emotion recognition accuracy (50%) while the VR DA technique does not bring any enhancement with the system achieving the same accuracy (53%). The HR and the TR DA techniques allow the FER system to slightly improve the performance by achieving an accuracy of 57% and 61%, respectively. The single DA technique that permits the FER system to reach the greatest accuracy, i.e., 75%, is the GAN. The combination of the two most performing geometric DA techniques, i.e., the HR and the TR, brings to a 10% increase of accuracy with respect to the no DA case (63% vs. 53%). However, the combination of these two techniques with the GAN allows to achieve the greatest overall emotion recognition accuracy, i.e., 83.2%, with a 30% increase when compared with the no DA case. I want to specify that the GAN generated images were not augmented with the geometric transformation, but just added to the set of geometric augmented normal images. Therefore, the size of the training DB when augmenting with the GAN & HR & TR combination was 2450. With regard to the recognition accuracy of single emotions, the sadness, happiness, disgust and neutral emotions were recognized with 100% accuracy whereas fear, surprise and anger achieved 75%, 60% and 50% accuracy, respectively.

1.4.2 Second experiment: Testing with the ExpW DB

In the second experiment, the system is tested with the ExpW DB, which contains a quantity of images much larger than the training DB (KDEF) and with more diverse face variations. The accuracy has been computed only for the case of no DA and for the combinations of DA techniques that achieved higher accuracy in the first experiment. It can be seen that the accuracy achieved by training only with the

Method	**Training DB**	**Accuracy**
Proposed	KDEF	83.20 %
da Silva *et al.* [dSP15]	MUG	45.60 %
da Silva *et al.* [dSP15]	JAFFE	48.20 %
da Silva *et al.* [dSP15]	BOSPHOROUS	57.60 %
Lekdioui *et al.* [LMR+17]	KDEF	78.85 %
Gu *et al.* [GXV+12]	JAFFE	54.05 %
Hasani *et al.* [HM17]	MMI+FERA	73.91 %
Mollahosseini *et al.* [MCM16]	6 DBs	64.20 %
Zavarez *et al.* [ZBO17]	6 DBs	**88.58 %**

Table 1.3: Comparison among state-of-the-art cross-database experiments tested on the CK+ DB.

original dataset is really low, i.e., 15.7%. The combination of HR with TR techniques allows to increase the accuracy up to 20%. However, also in this case, the greatest accuracy is obtained using a combination of GAN with HR and TR for augmenting the training dataset. The achieved accuracy is 43.2%, i.e., 27.5% higher than the case of no DA and comparable to the 30% increase obtained in the first experiment. This confirms that this combination of DA techniques is effective to improve the emotion recognition accuracy of FER systems. With regard to the recognition accuracy of single emotions, the neutral emotion achieved the greatest accuracy, i.e., 88.2%, followed by happiness (62.5%), surprise (56.2%), sadness (28%), anger (20.7%), and finally disgust and fear (both 20.3%). These results disagree with those obtained in the first experiment (except for neutral, happiness and surprise emotions) and the likely reason is the difference between the tested datasets.

1.4.3 Comparison with the state-of-the-art

Finally, in Table 1.3, the greatest accuracy achieved by the proposed FER system has been compared with those achieved by state-of-the-art cross-database experiments conducted for FER systems and tested on the CK+ DB. It can be seen that the approach proposed by [ZBO17] is the only one that outperforms our proposed approach. However, they trained their model using a bigger dataset composed by 6 DBs (more than 6,200 images). Furthermore, it must also be noted that our results are achieved by using only 2 images to generate novel synthetic training images with the GAN. Probably, by augmenting the training DB with additional images the performance will improve, which will be the objective for future experiments. Further details regarding the state-of-the-art studies considered for the comparison are provided in Section 1.2.

1.5 Conclusions

The use of GAN-based DA has been experimented and combined also with other geometric DA techniques, for FER purposes. The "augmented" KDEF DB was trained using the well-known VGG16 CNN neural network and the CK+ and ExpW DBs were used for testing. The results demonstrate that geometric image transformations, such as HR and TR, provide limited performance improvements; differently, the adoption of GAN enriches the training DB with novel and useful images that allow for quite significant accuracy improvements, up to 30% with respect to the case where DA is not used. These results confirm that FER systems need a huge amount of data for model training and foster the utilization of GAN to augment the training DB as a valid alternative to the lack of huge training DB. The drawback of GAN is the relevant computational complexity that introduces significant times for the generation of the synthetic images. Specifically, in our experiments, 3 days were needed to reach a loss value of 0.02 for training the GAN network, from which I obtained the augmented dataset that permitted to reach 83.20% of accuracy. This result highlights the relevance of training data for emotion recognition tasks and the importance of considering the correct DA technique for augmenting the training dataset when large datasets are not available. A possible solution to reduce the time needed to train the GAN network may be to generate a lower number of synthetic images with the GAN and combine this technique with less complex geometric DA techniques.

Part II

From FER to QoE

Chapter 2

QoE approach evaluation from Facial Expressions

2.1 Introduction

Since the PhD research main objective was evaluating the QoE automatically, my studies led my research to approach the QoE evaluation in a different way. As a first approach, I thought to find a correlation between the main 7 facial emotions and QoE values. Finding a correlation between only 7 emotions and the 5 scores of the Absolute Category Rating (ACR) scale that are indicative of the perceived QoE value was the first problem. Shaping this problem in an ML approach, it is possible to consider the 7 emotions as features and the 5 scores of the ACR scale as the target to predict. In this approach the usage of a features' number close to the number of labels leads to assume that there is a correlation almost one-to-one between the felt emotion and one of the levels of the ACR scale. Furthermore, the QoE is given by the whole experience, then use a single categorized emotion to describe a whole experience could be too reductive. Therefore, a more common approach is to generalize the problem using more features and mapping them to the target. For this reason, I went deeper into the facial emotions field, trying to discover what can describe better the facial expression. The FACS is an anatomically based framework for decoding all distinct recognizable facial movement. It splits facial expressions into individual components of muscle movement, called Action Units (AU) [EF78]. From [VKA^{+}11] we can appreciate that from AUs is also possible to infer the felt emotion. Then by using the FACS is possible to increase the number of features for an ML approach.

Moreover, using this approach I could focus on the objective of investigating the potentialities to estimate the QoE automatically and unobtrusively by acquiring a video of the face of the subject from which facial expression and gaze direction are continuously extracted. This avoids bothering the subjects with questions to collect opinions and feedback. Indeed, while convenient and effective, self-report is problematic because it is subject to biasing from factors not related to the stimulus, such as the interviewer reaction to the questions, the way the questions are answered, and the context (tests are typically conducted in laboratory). Moreover, surveys and interviews are time-consuming and may be invasive and annoying for the users.

Therefore, this research specifically focused on the estimation of perceived QoE for video streaming services. To this aim, two different experiments have been conducted: 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; 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. The facial AU 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. Three facial metrics to derive significance from these facial features are defined.

Furthermore , the impairment level (IL) feature has been defined. It has the objective to highlight the level of impairment introduced to the video by the combinations of video KQIs concerning 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 ML classifiers aimed at QoE estimation. Specifically, two QoE estimation models are defined: 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.

2.2 Background

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 [EDM+17].

With regard to speech quality assessment using EEG, the studies in [ASA+12] and [UMMVA19] investigated the potentials of event-related-potentials (ERP) analysis as a valid tool to indicate variation in quality perception. In particular, in [UMMVA19], 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 [AASM16] concluded that for longer sequences, low-quality conditions led

Reference	Stimuli	Distortion	Assessment	Psychophysiological measure
[ABSM18]	2D videos	Bandwidth, delay, video resolution	Subjective test	Face expression
[TDL+15]	2D videos	No	Subjective test	Face expression
[AWZ12]	Speech	Delay, bandwidth and packet loss	Subjective test	Acoustic feature
[ASA+12]	Speech	Signal-correlated noise	Subjective test	EEG
[UMMVA19]	Speech	Signal-correlated noise	Subjective test	EEG
[AASM16]	Audiovisual	Coding at low bitrate	Subjective test	EEG and EOG (electrooculogram)
[KHL+14]	2D and 3D videos	Coding at high quantization parameter	Subjective test	EEG
[ARM14]	2D videos	Coding at low bitrate	Subjective test	EEG and eyetracking
[RBLC16]	2D videos	Coding at high quantization parameter	GCD and gaze analysis-techniques	Eyetracking
[RLC17]	2D videos	Packet loss	Subjective test	Eyetracking
[EPCZ10]	2D videos	Packet loss	Subjective test	Eyetracking
[JNBJ08]	2D videos	Induced amusement and sadness emotions	Subjective test	Face expression and 15 physiological measures

Table 2.1: Comparison of reference studies based on psychophysiological measures.

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 [KHL+14]. In [ARM14], 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 [RBLC16] 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 [RLC17]. 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 [EPCZ10]. For this reason, gaze direction is often integrated into image and video quality metrics with the objective to further improve their quality prediction performance [BBE+09].

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 [MH19], creation of predictive NR metrics for video quality assessment [TVMSL17], encrypted video streaming [SCW+19], human activity recognition [MBL+15], video QoE in wireless cellular networks [MJ19, CDW+17] and QoE prediction for Voice over IP (VoIP) calls [CPTP15, CM20]. Deep learning models are used in [LCL+18] for multimedia QoE prediction, in [LMSS20] for adaptive video streaming and in [TVMF+17] for live video streaming. For passive gaming video streaming applications, NR video quality estimation based on ML was used by [BJGM19]. A review of predictive QoE management using ML in video streaming services is presented in [TPDL18].

The relevant techniques of sentiment analysis and Facial Expression Recognition (FER) have also been recently explored by following ML-based approaches. In

[LdASOS17], 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. [XH19] 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. [KKKL18] 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. [DT15] 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 [ZZC$^+$16], 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 [SA14]. With regard to voice, [AWZ12] 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 [JNBJ08], 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 [TDL$^+$15], 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 [ABSM18], 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

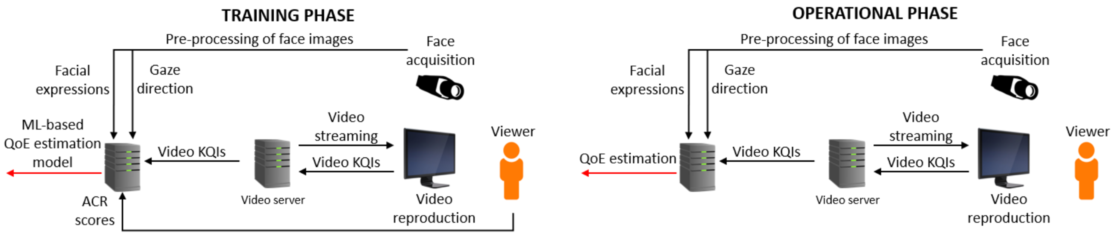


Figure 2.1: Proposed methodology: training phase and operational phase.

(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.

2.3 Proposed Methodology

The past studies, discussed in Section 2.2, 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 2.1, the framework of the proposed methodology is shown, 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 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.

It is to highlight that, concerning 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.

2.4 Data Collection

For dataset generation, two experiments were conducted to collect ground-truth quality perception values about video streaming services. One of these experiments, described in Section 2.4.1, was a crowdsourcing test in which participants had to watch and rate the quality of 20 videos impaired by long initial delays and rebuffering events. The other, described in Section 2.4.2, was a laboratory test in which participants were asked to watch and rate the quality of 105 videos subject to impairment caused by blurring.

2.4.1 Crowdsourcing test

The crowdsourcing test (Crowd) was carried out on the crowdsourcing platform Amazon Mechanical Turk (MTurk). By selecting 5 undistorted videos from the LIVE Mobile Stall Video Database[1] [GBY$^+$14, GBY$^+$16] 20 test video sequences were created, (codec: H.264, format: MP4, resolution: 1280×720 px, frame rate: 25 fps) introducing some long initial delays and buffering distortions. Specifically, the KQIs considered for this tests are the initial delay and the number and duration of buffering events. By combining these parameters 4 versions of test videos have been created:

- *Original (OR)*: 30-second version of the original video content without initial delay and buffering interruptions.

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

- *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, I 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) [ITU08]. 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 I was not able to actually verify those requirements from the platform. Therefore, I checked that these conditions were respected watching the videos of the recorded participants. I 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

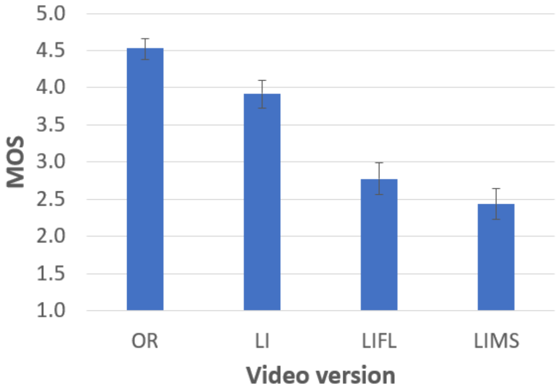


Figure 2.2: Crowd test: MOS with 95% CI for each video version.

for extracting facial expression features. Eventually, 20 participants completed the test successfully.

Outliers were identified through intra-rater and inter-rater reliability metrics, following the indications in [HSH+11, HKH+14, HHM+15]. 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. I defined as outliers the participants whose ratings obtained an intra-rater reliability lower than 0.75 [RGR+17]. As result I 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 2.2, the average MOS with 95% confidence interval (CI) computed for each test video version (OR, LI, LIFL and LIMS) is shown. The CI was computed using the Clopper-Pearson method as suggested in [HHVSK18] 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.

2.4.2 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) [SLU+15]. All selected videos were cut to be 20 seconds long. 7 versions of the 15 original videos by introducing different levels of blurring have been created, 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 2.4.1). The average intra-rater reliability computed

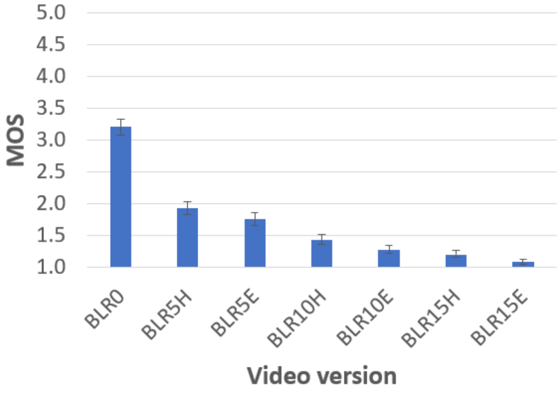


Figure 2.3: Lab test: MOS with 95% CI for each video version.

considering all 19 participants was 0.897. As for the Crowd test, I defined as outliers the participants whose ratings obtained an intra-rater reliability lower than 0.75. I 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 2.3, the MOS is showed with 95% CI computed for each video version. The CI was computed using the Clopper-Pearson method as suggested in [HHVSK18]. 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).

The present laboratory test aimed to analyze also the effects of visual quality degradation on perceived quality and emotional state of participants watching video clips. The emotional state has been controlled using a self-assessment manikin (SAM) method and by automatically analyzing the subjects facial reactions. The SAM questionnaire allows for highlighting if there is a correlation between the impairments of the video and the felt emotion. Thus, relationships between subjective measures of emotion and quality could be examined. The emotional state measurement has been collected by using the following approach: three discrete 9-class (valence, arousal and dominance) for the emotional dimensions evaluation from the self-assessment manikin (SAM) have been used [BL94]. Each scale title contained German antonyms: 'Wie körperlich entspannt / erregt sind Sie?' (English: 'How bodily relaxed / aroused are you?') for arousal, 'Wie emotional kontrolliert / unkontrolliert sind Sie?' ('How emotionally controlled / uncontrolled are you?') for dominance and 'Wie angenehm / unangenehm fühlen Sie sich?' ('How pleasant / unpleasant do you feel?') for valence.

2.5 Data Pre-processing

This section shows the procedures followed to pre-process the collected data, which are illustrated in Figure 2.4. 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 2.4. The proposed processing relies on: i) the 3 metrics that are defined in Section 2.5.1 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 2.5.2 that allows for comparing subjective scores collected from different tests; iii) the devised impairment level feature described in Section 2.5.3, which is aimed at making comparable the measured KQIs for streaming sessions with different types of impairments.

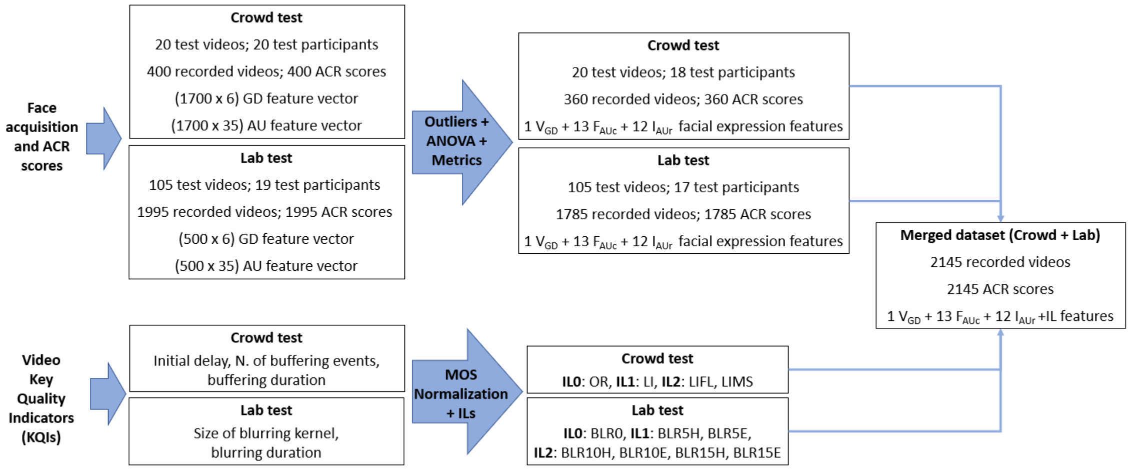

Face acquisition and ACR scores
Crowd test
20 test videos; 20 test participants
400 recorded videos; 400 ACR scores
(1700 x 6) GD feature vector
(1700 x 35) AU feature vector
Lab test
105 test videos; 19 test participants
1995 recorded videos; 1995 ACR scores
(500 x 6) GD feature vector
(500 x 35) AU feature vector
Outliers + ANOVA + Metrics
Crowd test
20 test videos; 18 test participants
360 recorded videos; 360 ACR scores
1 V_{GD} + 13 F_{AUc} + 12 I_{AUr} facial expression features
Lab test
105 test videos; 17 test participants
1785 recorded videos; 1785 ACR scores
1 V_{GD} + 13 F_{AUc} + 12 I_{AUr} facial expression features
Merged dataset (Crowd + Lab)
2145 recorded videos
2145 ACR scores
1 V_{GD} + 13 F_{AUc} + 12 I_{AUr} +IL features
Video Key Quality Indicators (KQIs)
Crowd test
Initial delay, N. of buffering events, buffering duration
Lab test
Size of blurring kernel, blurring duration
MOS Normalization + ILs
Crowd test
IL0: OR, IL1: LI, IL2: LIFL, LIMS
Lab test
IL0: BLR0, IL1: BLR5H, BLR5E,
IL2: BLR10H, BLR10E, BLR15H, BLR15E

Figure 2.4: Data pre-processing.

2.5.1 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, facial expressions and gaze direction features were extracted using the OpenFace toolkit [BZLM18, WBZ$^+$15, BMR15]. Specifically, the task has been based on the Facial Action Coding System (FACS) to analyze the facial expressions by means of facial Action Units (AUs) [EF78]. 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

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

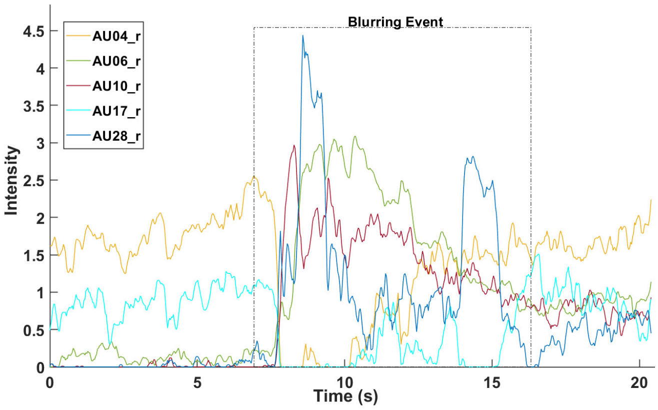


Figure 2.5: Temporal evolution of the intensity of some AUs for a participant during the reproduction of the BLR10H test video.

allows for setting the face tracking confidence, which computes the confidence value of the face tracker regarding the facial landmark detection. The confidence value has set to 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. OpenFace library has also the 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 2.5 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 2.6 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 2.4, I 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, the features has been analysed to focus on those that were the most significant. For this reason, I 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, a one-way ANOVA has been computed for both the Crowd and Lab tests, whose results are shown in Table 2.2. Note that the F value (variance between the means of the considered populations) and the p

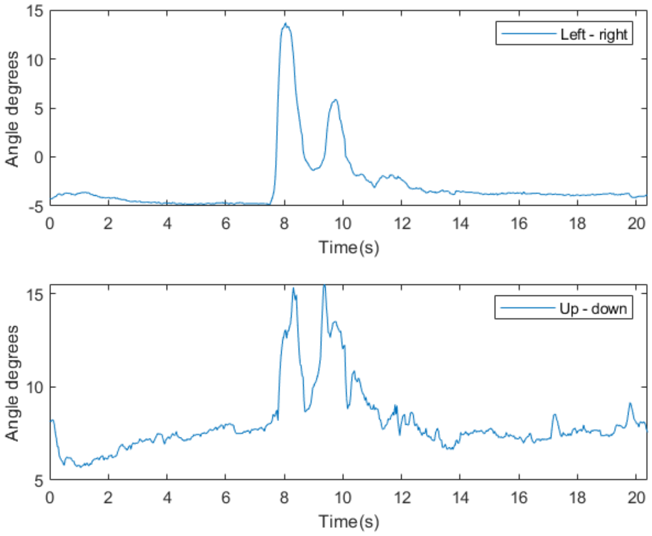


Figure 2.6: Temporal evolution of the gaze direction for a participant during the reproduction of the BLR10H test video.

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 this table are provided 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 [LCK$^+$10]. 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 2.2 are the most significant for QoE estimation and I computed 3 metrics from these feature arrays as explained in the following.

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	$>$ **0.001**	1.84E+14	< 0.001
$AU04_c$	2.16E+14	< 0.001	7.95E-01	$>$ **0.001**
$AU07_c$	1.32E+14	< 0.001	1.10E+14	$>$ **0.001**
$AU09_c$	1.57E+14	< 0.001	1.00E+14	$>$ **0.001**
$AU12_c$	1.05E+14	$>$ **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	$>$ **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	$>$ **0.001**
$AU01_r$	-1.50E+13	$>$ **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	$>$ **0.001**
$AU05_r$	1.33E+14	< 0.001	1.01E+14	$>$ **0.001**
$AU06_r$	1.30E+14	< 0.001	9.70E-01	$>$ **0.001**
$AU09_r$	1.31E+14	< 0.001	9.84E-01	$>$ **0.001**
$AU10_r$	1.40E+14	< 0.001	9.59E+04	< 0.001
$AU15_r$	1.42E+14	< 0.001	9.81E-01	$>$ **0.001**
$AU17_r$	1.35E+14	< 0.001	9.40E-01	$>$ **0.001**
$AU23_r$	1.23E+14	< 0.001	9.87E-01	$>$ **0.001**
$AU28_r$	1.26E+14	< 0.001	9.62E-01	$>$ **0.001**
$AU45_r$	1.34E+14	< 0.001	9.99E-01	$>$ **0.001**

Table 2.2: ANOVA results for the considered features.

The *frequency of activation* $F_{AU_c^j}$ of the j-th AU_c over the activation of all the AU_c was computed, as in Eq. (2.1). Activation means that the AU_c assumes a value greater than 0.

$$F_{AU_c^j} = \frac{\sum_{n=1}^{N} a_n^j}{\sum_{n=1}^{N} \sum_{j=1}^{J} a_n^j} \tag{2.1}$$

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, the *intensity of the activated* AU_r, $I_{AU_r^k}$ was computed, as in Eq. (2.2).

$$I_{AU_r^k} = \sum_{n=1}^{N} AU_{r,n}^k \tag{2.2}$$

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

Finally, a definition of the *variance of the GD*, V_{GD^g} is given, 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. (2.3).

$$V_{GD^g} = \frac{\sum_{n=1}^{N} (GD_n^g - \overline{GD^g})}{N} \tag{2.3}$$

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, I was 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, I 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 2.6.3 are: 1 V_{GD}, 13 F_{AU_c} and 12 I_{AU_r}. It is important to highlight that by computing these metrics I obtain the same number of features for each recorded video, regardless the video length. This is important to make feature datasets compatible for merging.

2.5.2 ACR scores

As illustrated in Figure 2.4 and discussed in Sections 2.4.1 and 2.4.2, 400 and 1995 ACR scores have been collected for the Crowd and Lab test, respectively. After outliers removal, I kept a total of 360 and 1785 ACR scores, respectively.

As it can be noticed in Figures 2.2 and 2.3, participants of different tests used the MOS scale differently for rating the video quality. Indeed, blurring impairments

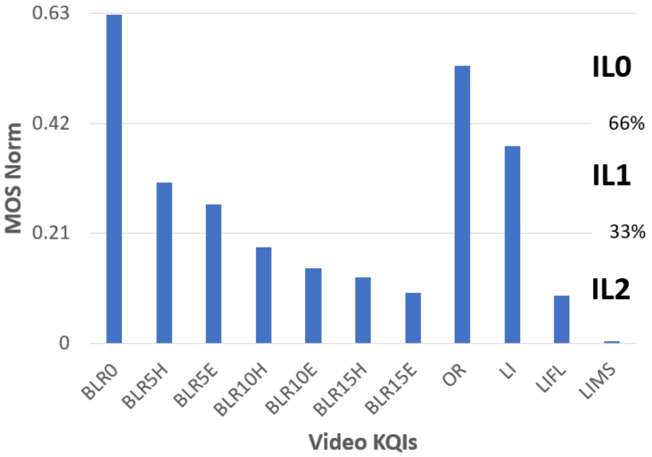


Figure 2.7: Normalized MOS computed for both the datasets and definition of the ILs.

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, the ACR scores were normalized applying for both the datasets the mean normalization method as follows:

$$ACR_{norm} = \frac{ACR - ACR_{mean}}{ACR_{max} - ACR_{min}} \tag{2.4}$$

where $ACR_{mean} = MOS$, $ACR_{max} = 5$ and $ACR_{min} = 1$. ACR_{norm} values are mapped in the [-1;1] range. Then the normalized MOS from the obtained normalized ACR scores has been computed, which is shown in Figure 2.7 (for easier comparison, MOS values were also shifted to set the lowest value to 0), from which I define the impairment levels as described in the next section. Then the ACR score datasets were merged together (and corresponding features) for training the ML-based QoE estimation model presented in Section 2.6.3.

2.5.3 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 2.5.1, has been considered 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, the IL feature has been defined, 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 I relied on the normalized MOS shown in Figure 2.7 and I decided to consider 3 ILs. I 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. I 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 I considered for training the ML-based model. I 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.

2.5.4 Emotional state analysis

To evaluate the correlation between the human emotions, visual impairments and perceived quality, a data analysis approach was used: The effects of experimental quality manipulations were examined by means of *multivariate analysis of variance (MANOVA)*. Hereby, participants' subjective reports could be tested for significant underlying relationships between visual degradation factors, perceived quality and emotional responses. More closely, this approach mirrors analyses of explicit user feedback typically conducted in the field. Two repeated-measures MANOVAs were calculated with ACR, valence, arousal and dominance ratings as well as correct decision as dependent variables and either *blurring level* (0, 5, 10, 15) or *blurring length* (no, half, whole) as within-subject factor. In case of statistically significant effects, pairwise comparisons were computed only between the degraded levels of *blurring level* (5, 10, 15) and *blurring length* (half, whole) using multiple paired t-tests with Holm correction. The statistical significance levels for the MANOVA and post-hoc analysis were both set to $\alpha = 0.05$ and in the former case Šidák-adjusted ($\alpha_{SID} = 0.0253$). Multivariate analysis suggested significant effects of *blurring level* on ACR rating ($F[3] = 192.79$, $p < 0.001$), arousal rating ($F[3] = 4.075$, $p = 0.011$), valence rating ($F[3] = 7.06$, $p < 0.001$) and correct decision ($F[3] = 18.077$, $p < 0.001$). In post-hoc analysis, all pairwise comparisons between blurring levels with regard to ACR, arousal and valence ratings turned out significant (all $p < 0.001$). For correct decision, the lowest blurring level significantly differed from the two higher ones (5 vs. 10, $p < 0.01$, 5 vs. 15, $p < 0.001$). Furthermore results manifested significant effects of *blurring length* on ACR rating ($F[2] = 269.1$, $p < 0.001$), valence rating ($F[2] = 7.057$, $p < 0.01$) and correct decision ($F[2] = 13.228$, $p < 0.001$). Post-hoc pairwise comparisons between half- and whole-blurring conditions were only significant for ACR rating ($p < 0.001$). Line plots in figure 2.8 contain mean values for each factor (blurring length: plot 1–3, blurring level: 4–6). The mean arousal rating increased across blurring intensities (from 0 to 5 to 10 to 15): From 4.088 (standard error of the mean: 0.11) to 4.33 (0.082) to 4.549 (0.092) to 4.711 (0.1).

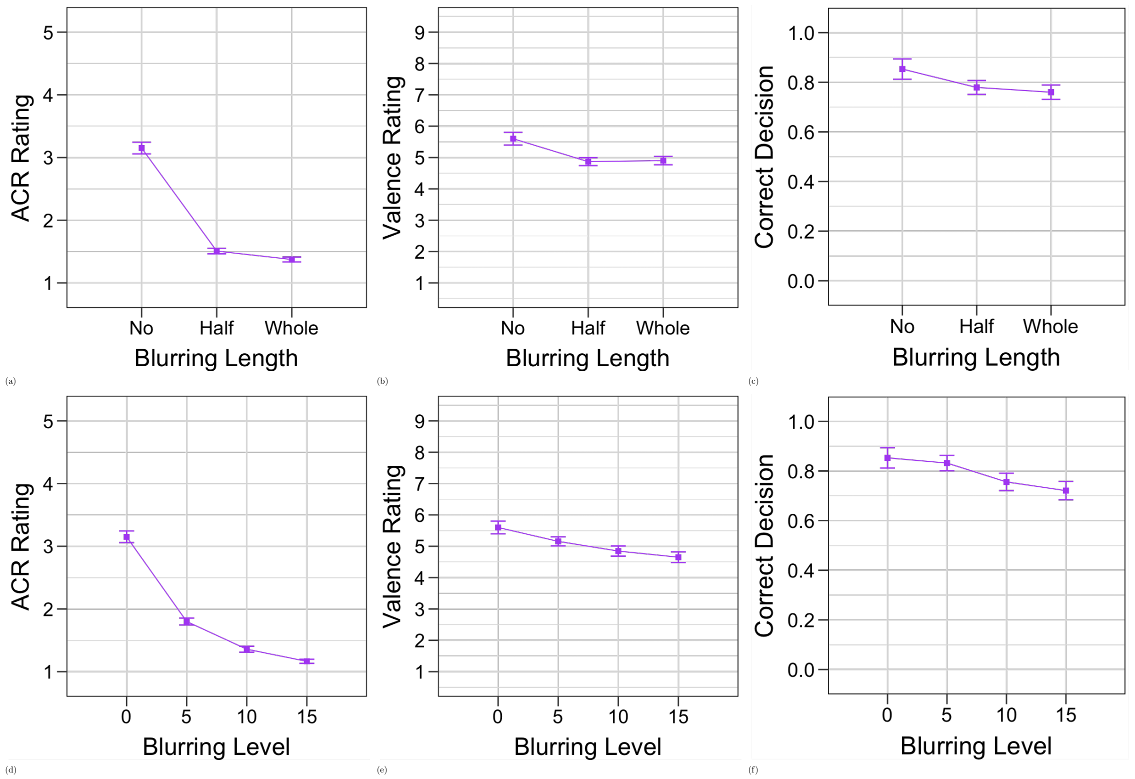


Figure 2.8: Effects of experimental factors *blurring length* (top row) and *blurring level* (bottom row) on ACR rating (left column), valence rating (middle column) and correct decision (right column). Purple squares represent arithmetic means, error bands represent 95% confidence intervals ($N = 19$).

2.6 ML-based QoE Model

In this section the considered features, (Section 2.6.1), the considered DA approach (Section 2.6.2) and the proposed QoE estimation models have been described(Section 2.6.3).

2.6.1 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 this 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 2.5 and illustrated in Figure 2.4.

2.6.2 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), a class over-sampling method for creating synthetic samples instead of replicating samples in the dataset was used. Specifically, the adaptive synthetic (ADASYN) algorithm which balances binary classes was used for the present study [HBGL08]. 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 [HYGS08]. 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. As explained above, the majority class has always the same dimension. The other classes were augmented to reduce the gap between the classes.

2.6.3 QoE estimation model

Two ML-based QoE estimation models based on the features described in Section 2.6.1 were defined. 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 I considered the three quality scales in Table 2.3. The best performance results were obtained with the k-NN classifier in all cases.

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		

Table 2.3: The three considered quality scales.

2.7 Experiment Results

In this section, the performance of the ML-based QoE estimation models proposed in Section 2.6.3 is evaluated. It has been done using three performance metrics for ML, i.e., accuracy, sensitivity, and specificity.

Firstly, the model performance is compared with different training/validation rates (90%/10%, 80%/20%, and 70%/30%) and k-fold cross-validation ($k = 5$, and $k = 10$). For this comparison I used the AU&GD&KQItoQoE model implemented with the k-NN classifier using the original Crowd+Lab dataset without DA. Results are provided in Table 2.4. 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, the specificity and sensitivity metrics are analysed. 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'. Then, this combination is considered 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 DA. Results are shown in Table 2.5. It can be seen that considering the $KQIIL_l$ feature the accuracy in-

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

creases 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, the model has been trained also 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.
Finally, in Table 2.6 I 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 I performed two cross-dataset evaluations (applying DA on the training dataset) and an evaluation using the combined (Crowd+Lab) augmented dataset (see Section 2.6.2) and a 70%/30% training/validation combination with $k = 5$ (best combination obtained in Table 2.4). 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 2.5. This suggests that the impairment level defined in Section 2.5.3 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 I am 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 reproduction. For example, a 3-level quality scale is used in [LKB+19] to evaluate whether the video quality is acceptable or annoying (Not Annoying / Annoying but acceptable / Not acceptable).

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

Table 2.4: 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 DA.

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%

Table 2.6: 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	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%

Table 2.5: 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 DA.

2.7.1 Comparison with state-of-the-art QoE models

In this section I 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, I 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 author knowledge, the only valuable ML-based model based on facial expressions for QoE estimation is [ABSM18]. Indeed, the model in [TDL$^+$15] 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 [ABSM18] 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, I compare the performance of our model with those of QoE objective models for the single tests. With regard to the Lab test, I used the NR perceptual blur metric defined in [CDLN07] and implemented with MATLAB[Bao20] 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, I 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, the ITU-T P.1203 model [ITU17] implemented with Python has been used used [RGR$^+$17, RGR$^+$18][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. I set the model to the *Mode 3* to access the full bitstream and I 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

2.7.2 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 2.3. However, in this work, these procedures have been implemented offline in our lab server and I 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, the impact of the face acquisition frequency on the QoE estimation is discussed, 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, specific experiments using the Lab dataset have been conducted .

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

Table 2.7: Impact of face acquisition frequency.

I 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. I 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). Six feature matrices (as defined in Section 2.6.1 but without the KQI feature) have been obtained, one for each of the 6 considered sample rates. Then the PCC between these feature matrices and the ideal (the one without sampling) feature matrix has been computed. Table 2.7 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, I 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.

2.8 Conclusion

In this work, I 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, I 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 I achieved, and an RMSE of 0.55, which is five times the one I obtained. I 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 I 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.

Part III

Smart context applications

Chapter 3

Approaching QoE to smart scenarios

The previous sections highlight that a correlation between expressions, emotions and perceived quality exists. Based on the obtained good results of the application in a smart scenario, as the next step, I decided to test this approach in a smart context. For this kind of approach, the typical scenario could be a smart city or a smart room, in which people can be monitored with the previous techniques. For this reason, smart-cities and smart-rooms environments have been studied. At the moment that I started studying them, there was not any declaration of QoE about these two environments. Thus, I started defining a smart-city QoE modelling approach.

3.1 Quality of Experience Management of Smart City services

The increasing efforts in research activities related to the Quality of Experience (QoE) reflect the fact that perceived quality is a key criterion for evaluating systems, services or applications during the design phase or during operation [MR14]. Traditionally, the QoE is mainly focused on networked multimedia and communication applications, where the person that uses an ICT product takes the role of a user [LCMP12]. However, authors of [MR14] extended the QoE concept to a more global view providing the following definition: "*QoE is the degree of delight or annoyance of a person whose experiencing involves an application, service, or system*". The inclusion of the terms 'system' and 'person' allows for considering the QoE even for those experiences where the interaction with the application, service or system is not at the core of the consideration, e.g., a person attending a concert. Also, QoE is defined as "*an assessment of the human experience when interacting with technology and business entities in a particular context*" [LC12].

In this work, I aim to discuss the applicability of the aforementioned QoE concepts to Smart City services. Smart Cities aim to create sustainable economic development and high quality of life by excelling in multiple key areas such as government, transportation, environment, healthcare, living, energy. ICT infrastructure and services represent one of the enabling factors for the actual implementation of Smart Cities [BM15]. In a city *traditional public services* are those services offered to the citizens, such as public transport, healthcare, education, for which the presence of ICT was limited or not necessary. *Smart City services* may then be considered as an 'extension' or an 'evolution' of *traditional public services*, supported by an important (but not essential) presence of ICT systems, with the objective to make everyday human activities better supported by digital systems. As the offered smart public services have to be user-centric, the key challenge is to evaluate, design and standardize new solutions oriented not only to ensure high performance with respect to the Quality of Service (QoS) but also to guarantee high levels of QoE perceived by the citizens.

To the best of author knowledge, the study in [BM15] is the first and only one discussing QoE for Smart Cities. The authors provide an initial analysis and indicate some factors related to QoE in the smart city context, i.e., usability, personalization, usefulness, transparency, accessibility, efficiency, learnability and findable. However, how QoE should be managed in the Smart City context is not discussed, which is the aim of this work.

3.1.1 Quality management for traditional services

Traditional public services refer to those services offered to the citizens, such as public transport, healthcare, education, for which the presence of ICT was limited or not necessary. For these services, most of quality models proposed in the literature measure the service quality as the difference between the customer's expected service quality and perceived service quality, along different quality dimensions (QDs) [Yar14]. Expected and perceived service quality are collected from customers through customers' interviews. One of the most relevant models is SERVQUAL, a multidimensional quality model for measuring customers' perceptions as a function of 5 service QDs (assurance, empathy, reliability, responsiveness and tangibility) that can be generalized to any type of service [PZB88]. Indeed, SERVQUAL has been used as the basis to measure the quality of several traditional public services, such as student's satisfaction [HIRR09], healthcare service quality in a hospital [Lee17] and quality of transit services [BDF16].

Fig. 3.1 shows the quality management framework for traditional public services. Surveys and interviews are conducted to collect expected and perceived quality from customers as well as the importance of the considered QDs for that specific service. This is useful to understand which QDs have more influence on the quality perceived by the user so that quality management actions may be directed to specific QDs. Objective measurements are also conducted by the service providers to evaluate the service performance, which should satisfy well-defined targeted quality values decided by the service provider during service planning, and can be seen as the QoS in

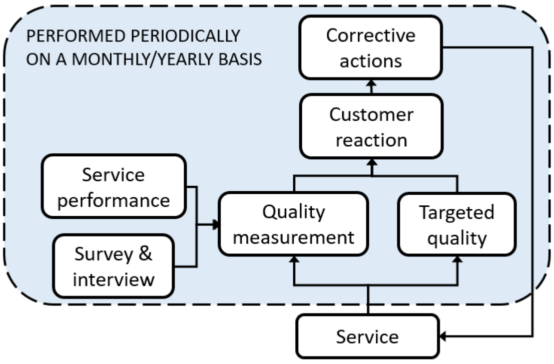


Figure 3.1: Quality management framework for traditional public services.

telecommunications systems. Then, the comparison of objective measurements with the collected subjective opinions is aimed at investigating whether the customers are satisfied with the provided service and, on the contrary, to suggest the QD of the service that are not appreciated by the customers and need intervention.

In the following list the limitations with reference to quality management for traditional services are highlighted:

1. *Quality monitoring is not automatic*: service providers conduct time consuming and costly quality measurements, mainly through surveys and interviews.

2. *Influence of QD variation is not investigated*: quality measurements regard different QDs but how the variations of these QDs may influence the overall perceived customer quality is not investigated.

3. *Objective and subjective quality measurements are collected at different times*: objective and subjective quality measurements should be measured simultaneously to find a correlation between them.

4. *QoE models cannot be defined*: as a consequence of points 2) and 3), no QoS-to-QoE model can be derived from quality measurements.

5. *Difficulty to adjust quality in real-time*: even if a model is defined, generally it is not possible to 'act in real-time' to adjust the quality. As an example, consider transport service: if a line becomes crowded, the service provider cannot instantly decide to provide more vehicles to serve all customers; generally, the number of customers for each line are observed for a certain period of time to decide how to distribute the vehicles to provide the best service to the customers.

6. *Long temporal scale*: quality evaluations and resulting correcting actions for quality improvement are performed on a long temporal scale (monthly/yearly) making this process inefficient in case quality degradation events occur between two quality evaluation times.

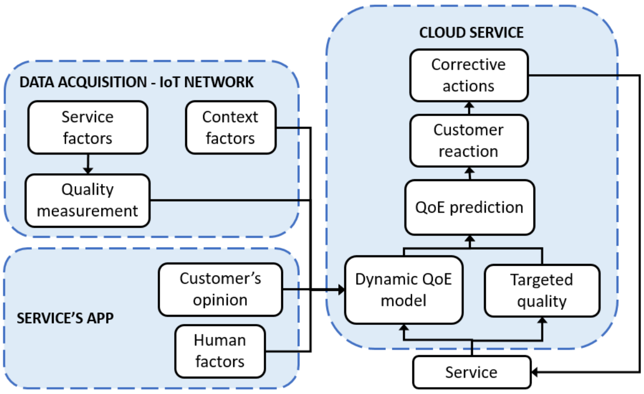


Figure 3.2: Proposed QoE-aware management framework for smart services.

3.1.2 Quality management for Smart City services

Smart City term is referred to those traditional public services which are importantly (but not essentially) supported by ICT for providing enhanced service quality and supplementary functionalities to the customers. For example, in smart healthcare and smart transportation the essential elements of the service remain respectively the doctors and the means of transport; however, ICT may improve overall customer perceived quality by providing digital and additional functionalities to the customers such as telemedicine for healthcare and smart ticketing solutions for transport services.

As defined in [LCMP12], QoE is influenced by system, human and context factors. For pure ICT services (e.g., video streaming, VoIP), the QoE is typically evaluated by conducting subjective quality assessment in laboratory, i.e., standardized and controlled tests in which people are asked to rate the perceived quality for a specific service. Quality ratings provided by users are used as ground-truth to derive a correlation between variations of influence factors and variations of perceived QoE, which are often nonlinear [RESD10, FHTG10b, TLPM17]. Mathematical QoE models are then built 'off-line' and can be used 'on-line' to manage almost in real-time the service's resources to provide a satisfactory service to the users. Unfortunately, it is not possible to replicate in a laboratory a Smart City service and study how QoE is influenced by influence factors. Therefore, the QoE must be estimated *on the fly* while the service is provided. Also, besides human, system and context influence factors, for QoE evaluation of Smart City services I also consider service influence factors, which are those factors that have an important impact on the perceived service quality. For example, influence factors for transport services are vehicle cleanliness, space on board, waiting time [BDF16].

In Fig. 3.2, it is shown the proposed QoE-aware management framework for Smart City services. Besides allowing more automated and digitally assisted services, ICT systems may also be used to support the evaluation and management of the quality of smart services by supplementing limitations of quality management for traditional public services. The technologies I consider for achieving this objective are the Internet of Things (IoT) and machine learning (ML). Sensors and actuators taking part in an IoT network can be used to acquire and collect objective quality measures of service performance. For example, with regard to the transport service, the GPS may be used to track the vehicle position and estimate bus frequency and regularity; also, passenger's comfort may be measured in terms of space on board, vehicle cleanliness and driving style. The first two can be estimated with ML algorithms from images recorded by cameras installed inside the vehicle. The driving style can be estimated by means of sensors (e.g., accelerometer) installed inside the vehicle. In this way, most of quality measures can be collected without asking customers about the perceived quality, avoiding to annoy them and saving time and money. Moreover, sensors may be used to measure context factors such as environment factors, time of day, mobility.

A service's App may be developed to provide the customers with a multimedia human-to-machine interface with the service, which can be used to select service settings and preferences as well as to visualize service information. Also, by means of this App, the customer will be able to provide feedback regarding the perceived service quality. Considering again the transport service as an example, the App may provide the customer with supplementary functionalities such as find the vehicle stop and read the timetables as well as buy and validate electronic tickets. When the trip is over, the customer may leave comments about the perceived quality of the trip. This App feature will replace the traditional manual conduction of survey and interviews for collecting customer's opinions. Furthermore, having a personal account allows the collection of human factors such as age, sex.

Quality estimation and management will run in a cloud system. The objective measures of system performance obtained through the IoT devices are used in QoE models and can be complemented with subjective opinions that are collected during and after the provisioning of the service through the service's App, which clearly involve directly the users. Therefore, in this case the correlation of objective and subjective data would be possible as these measurements are collected automatically and simultaneously thanks to the ICT elements. Continuous data observation allows to build dynamic QoE models, which are a function of system, human, context and service factors and evolve with the time as a function of changes in the observed data. Also, different and personalized QoE models can be created, which may differ on the basis of the customer's profile.

The predicted QoE has an impact on the customer reaction (service dissatisfaction, leaving the service, not paying, influencing the opinion of other people also through social networks). As a consequence, the provider takes some correcting actions that impact mostly on the resources, prices and communications activities. These clearly have an impact on the users and are discovered through measurements. The support of ICT allows to constantly monitor the quality of the provided service and to promptly react each time insufficient quality is observed.

The challenges raised by the proposed QoE-aware management framework are not trivial. First, different Smart City services have different requirements and provide different functionalities. Second, most of ICT technologies which may support smart services are emerging and not standardized, such as the IoT, Wireless Sensor Networks (WSN) and cloud services. Interoperability and scalability issues should be considered as well as data privacy. Therefore, the measurement and management of the QoE for these services and technologies should be accurately designed and tested for different and practical scenarios.

3.1.3 Conclusion

This work investigated the applicability of QoE management on Smart City services. I discussed the limitations of quality management for traditional public services (services for which the presence of ICT technologies is limited or not necessary). Then, I proposed a potential framework for the QoE management in Smart City services (traditional public services supported by an important, but not essential, presence of ICT systems), which is based on Internet of Things and machine learning for QoE prediction and management.

3.2 Ambient Illumination and Noise Effects on QoE

Investigating the smart city scenario, I concluded that the environment itself could be an impact factor for the perceived quality. Control the QoE level in a wild scenario without having the managing of the external stimuli and also without knowing which stimuli could appear, it is not the best scenario to make a study about a QoE management system. For this reason, in the next sessions, I set a laboratory test, in order to understand first if the environment could affect the perceived QoE and if the emotion could help to understand the impact of different ambient condition illumination and noise sounds. As a first step, the objective of this study is to evaluate the effects of ambient illumination and noise on two multimedia consumption scenarios: watching a video on TV and reading a comic strip on a tablet. To this aim, I organized an experiment furnishing a room with a sofa, a TV and a tablet to implement the video streaming and reading sessions in a simulated home environment. Different combinations of ambient illumination were simulated alternating bright light ambient, soft light effect and dark condition. Furthermore, an annoying noise to simulate an external noise disturbing the multimedia consumption of the participants was selected. Then, I conducted a subjective quality assessment involving 20 people, which were asked to rate the perceived QoE using the Absolute Category Rating (ACR) 5-level quality scale and to express their emotions completing the Self-Assessment Manikin (SAM) questionnaire. Finally, the impact of illumination and noise on ACR ratings and SAM scores is evaluated computing the Multivariate

Analysis of Variance (MANOVA). A linear parametric model to predict the QoE based on illumination and noise context factors has also been proposed.

3.2.1 Related Works

Most of the studies regarding the evaluation of the QoE for multimedia services investigate the effects of impairments affecting the network state, the media quality and the application performance [FHTG10a, PBC16, SES$^+$15]. Some studies also took into account the influence of the context on QoE. [MHSK$^+$18] focuses on the potential of enhancing QoE management mechanisms by exploiting valuable context information. [LHM$^+$18] investigates the value of context-awareness in bandwidth-challenging HTTP adaptive streaming scenarios. [SNR15] proposes a proactive adaptation policy to prevent application unusable state due to contextual changes. These states are classified as: computing environment (CPU, memory, available networks), user environment (physical and mental state of the user) and physical environment (location, weather, noise). In [MZ5], the authors propose, develop and validate a Context-aware approach for Quality of Experience modelling, Measurement and prediction (CaQoEM), which incorporates several context attributes and QoE parameters to measure and predict users' QoE under uncertainty using Bayesian networks, utility theory and a bipolar scale.

To the best of author knowledge, [WSSE15] is the only study considering the ambient noise as a factor influencing the QoE. Specifically, they focused on the users of a mobile video streaming system that can not appreciate the sound of a video if the volume of the ambient noise is too high. They propose to monitor the ambient noise with the microphone of the mobile device and to increase automatically the volume of the audio track in case the ambient noise exceeds a threshold of 75 dB. Moreover, the effects of ambient illumination on QoE have been considered for assessing 3D video quality and multimedia services. In [NDKAK12], the effects of ambient illumination on 3D video quality and depth perception are investigated by conducting subjective tests under four different ambient illumination conditions (i.e., 5, 52, 116, and 192 lux); 5 lux corresponds to a dark condition, while 192 lux indicates a bright light environment. From the results, it has been observed that when the ambient illumination in the content access and consumption environment increases, the MOS ratings of the viewers for the perceived video quality also increase. Conversely, when the ambient illumination increases, the perceived depth MOS ratings of the viewers decrease. In [JAPM18], a subjective quality assessment is conducted where participants watched 10 video sequences at different resolutions and bitrates enriched with light, vibration and air flow sensory effects. With regard to the lights, the results indicate that additional light effects reduce eye strain due to a smoother lighting difference between display and background.

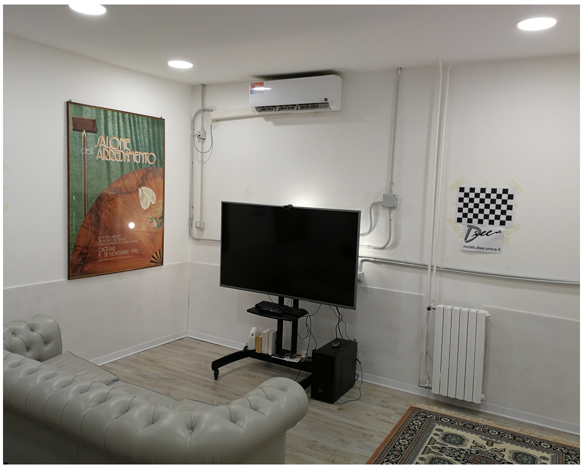


Figure 3.3: Room where the experiment was conducted.

3.2.2 Methodology

In this Section, the considered scenario and the details regarding content preparation, experiment design and subjective quality assessment have been described.

Considered Scenario

The objective of this research is to investigate the effects of ambient illumination and noise on the quality perception related to two different multimedia consumption scenarios: watching a video on TV and reading a comic strip on tablet. To this aim, I organized an experiment and conducted a subjective quality assessment for quality evaluation. I furnished a room with a sofa, a TV and a tablet to implement the video watching experiment and the reading experiment as if they were consumed in a home environment. In Fig. 3.3, I show a picture of the room, which is equipped with 6 overhead lights (OL) and 3 dome lights (DL) behind the TV. I set the experiment focusing on different combinations of room illumination and presence or absence of annoying noise to introduce ambient context stimuli to the video watching and reading experiences. For the video watching experiment, 3 different illumination conditions have been considered: i) overhead lights on and dome lights off (bright light ambient); ii) overhead lights off and dome lights on (soft light effect); iii) overhead lights off and dome lights off (dark condition). I did not consider condition iii) for the reading experiment as with all lights off it was too dark to read the comic strip on the tablet, which was set on reading mode to simulate reading on paper. All the illumination conditions were implemented with both the presence and absence of annoying noise. The audio file used to introduce the annoying noise was downloaded from YouTube and is the audio of a sander with sound intensity of 110 dB. I cut a 13-second audio sequence from the original audio. For the experiment conditions considering the presence of noise, the audio of the sander is reproduced in the room

2 s later the start of the watching/reading experience.

Content preparation

For the video watching experiment, 6 different video contents belonging to the same genre (action) were selected. The videos have a native resolution of 4K (3840×2160), were downloaded from YouTube, and do not have any kind of impairments. a 15-second video sequence from the original videos has been cut, so that all test videos had the same length. The cut videos were then encoded at the 4K resolution (3840×2160) at 30.00 fps with the MP4 codec.

The videos were reproduced on the SAMSUNG TV UHD 4K Flat Smart JU6800 Series 6 with a 60-inch diagonal screen and aspect ratio 16:9. The native screen resolution is 3840×2160 pixels. The Recommendation ITU-R BT.2022 [ITU12] suggests to watch a screen with 3840×2160 resolution at the optimal distance of $1.6H$, where H is the height of the TV screen. The SAMSUNG TV is 74.72 cm high, then the sofa where the participants sit to watch the TV was at the optimal distance of 1.2 m.

For the reading experiment, I selected 4 different Peanuts comic strips with the same structure, i.e., divided in 4 blocks with a gag in the last block. The comic strips are JPEG images with 1600×1381 pixels resolution. The comic strips were watched on the HUAWEI MediaPad M3 tablet, which has an IPS 8.4-inch diagonal screen with a WQXGA resolution, 2560×1600 pixels. The tablet was set in reading mode with the *Reading Mode* App for Android with the following settings: brightness 50%, BlueLight filter enabled, BlueLight filter density 10%, and Redness 10%.

3.2.3 Experiment design

The ambient stimuli (illumination and noise) are the independent variables of the experiment, whereas the results of the subjective assessment are the dependent variables.

Independent Variables

Ambient illumination and noise are the ambient stimuli affecting the watching and reading experience. I considered three ambient illumination conditions: i) OL off and DL off (only for the video watching experiment); ii) OL on and DL off; iii) OL off and DL on. I considered two ambient noise conditions: i) absence of noise; ii) presence of annoying noise in the room. Table 3.3 summarizes all the considered experiment conditions (ECs).

Dependent Variables

Absolute category rating (ACR) ratings and self-assessment manikin (SAM) scores are employed to capture respectively variation in perceived QoE and emotional state

EC	Media	End device	Illumination	Noise
1	4K video	TV	OL OFF & DL OFF	No
2	4K video	TV	OL ON & DL OFF	No
3	4K video	TV	OL OFF & DL ON	No
4	4K video	TV	OL OFF & DL OFF	Yes
5	4K video	TV	OL ON & DL OFF	Yes
6	4K video	TV	OL OFF & DL ON	Yes
7	Comic strip	Tablet	OL ON & DL OFF	No
8	Comic strip	Tablet	OL OFF & DL ON	No
9	Comic strip	Tablet	OL ON & DL OFF	Yes
10	Comic strip	Tablet	OL OFF & DL ON	Yes

Table 3.1: Experiment conditions (EC). OL: overhead lights; DL: dome lights.

of participants.

- A single discrete, five-class ACR scale with five category labels (Bad, Poor, Fair, Good and Excellent) according to ITU-R Rec. P.800 [ITU08];
- Three discrete, nine-class graphical rating scales for emotional dimensions (valence, arousal and dominance) from the SAM technique [BL94], which is a non-verbal pictorial assessment to evaluate emotions.

3.2.4 Subjective quality assessment

In total, 20 Italian participants, 5 females and 15 males between 25-45 years old, were recruited for the subjective quality assessment. All subjects reported normal or corrected-to-normal vision and were compensated at the end of the experiment with 5. The assessment was conducted in the test room shown in Fig. 3.3, which is located in the Department of Electrical and Electronic Engineering (DIEE) of the University of Cagliari, in Cagliari, Italy.

Before conducting the assessment, the participants were explained about all the different experiment conditions they had to evaluate. They were informed that the ambient conditions would be changed and they were asked to evaluate the overall perceived QoE by considering both media quality and ambient conditions. To give participants an overview of the different stimuli involved in the experiment, two training videos and comic strips were shown while the different conditions of room illumination and the presence of annoying noise were introduced. The participants were also trained about the utilization of the ACR quality scale for rating the overall QoE and the SAM questionnaire to express their emotions.

An ad-hoc platform has been developed to show the videos and comic strips to the participants as well as to collect the participants' QoE ratings and SAM scores. The platform has been developed using the Bootstrap open source toolkit for developing

Web applications: HTML5 was used for the front-end side and JavaScript for the back-end side. The WebRTC JavaScript library has been used for video streaming. The videos were played in full screen mode on the TV screen. Special attention has been given to the end user interface to create a user friendly platform. In fact, the platform guides the user through the video playing, comic strip reading, and rating phases with simple and intuitive buttons/indications. The participant interacts with the TV screen with a mouse (the keyboard is provided only for inserting the name) and with the tablet using its touch-screen. After watching each video (reading each comic strip) the participants were asked to rate the overall perceived QoE (including the impact of the ambient stimuli) and complete the SAM questionnaire on the basis of the perceived quality and emotions. After the submission of the rating scores, the next video (comic strip) was shown. The overall time needed to complete both the watching and reading assessments was about 15 minutes. It is important to highlight that the ECs in Table 3.3 were presented in random order to the participants.

3.2.5 Results

In Section 3.3.3, the results of the subjective quality assessment in terms of Mean Opininon Score (MOS) are presented. The impact of the independent variables (illumination and noise) on ACR ratings and SAM scores is evaluated based on the Multivariate Analysis of Variance (MANOVA) in Section 3.3.3. Finally, in Section 3.2.5, aa QoE prediction model to estimate the QoE based on the ambient stimuli is proposed.

Mean Opinion Score

In Fig. 3.8, I show the MOS scores with 95% confidence interval (CI) computed for each of the 10 ECs in Table 3.3. From these results, it can be seen that the noise had a really negative impact on the perceived users' QoE, especially for the video watching experience. After participating to the experiment, the participants commented that the noise was so annoying while watching the video that they were not able to follow what was happening on the video. However, some participants were able to concentrate and read the comic strip even with the presence of the noise. For this reason, the MOS scores for the reading experience disturbed by the noise (MOS between 2 and 2.5) are slightly greater than the MOS scores for the video streaming experience (MOS lower than 2). Moreover, the ambient illumination does not have any influence when the noise is present because this has total predominance on the perceived quality. On the other hand, when the noise is absent, the participants perceived a good QoE (MOS equal or greater than 4). Specifically, slightly better QoE is perceived when DLs are on and OLs are off, for both the video watching and reading experiences.

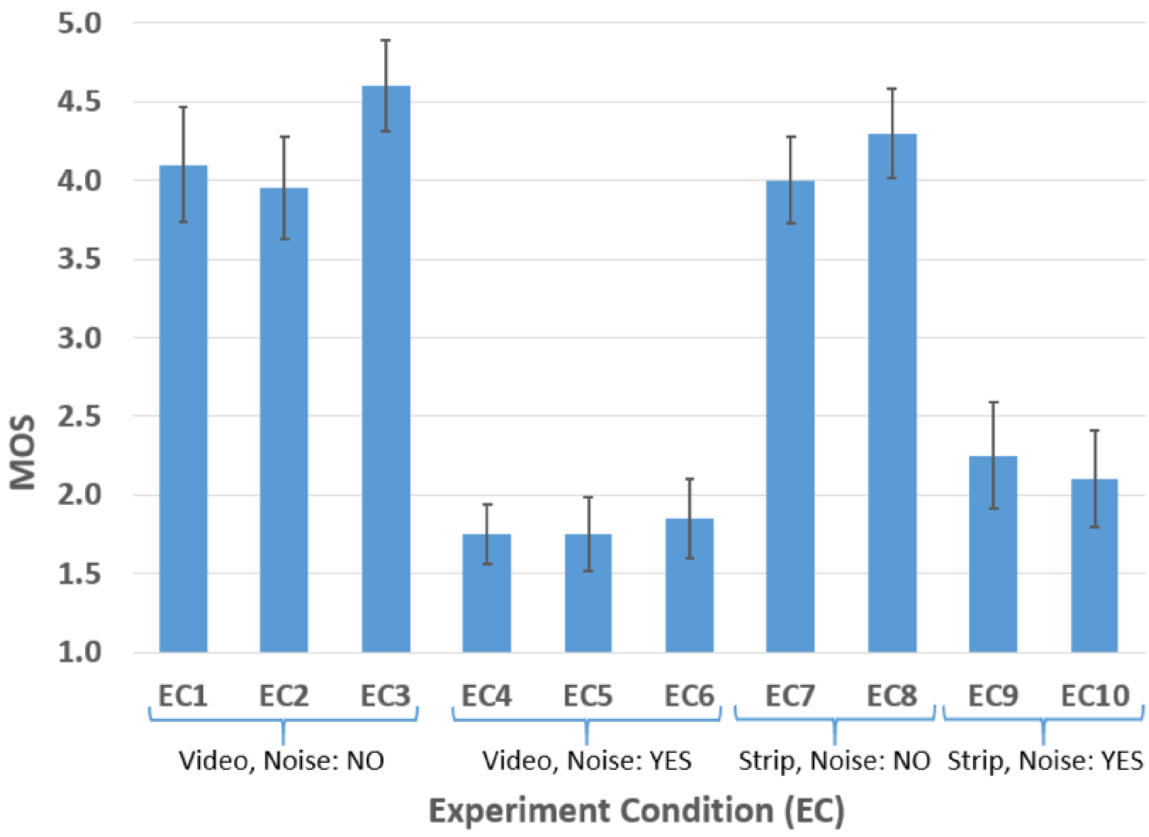
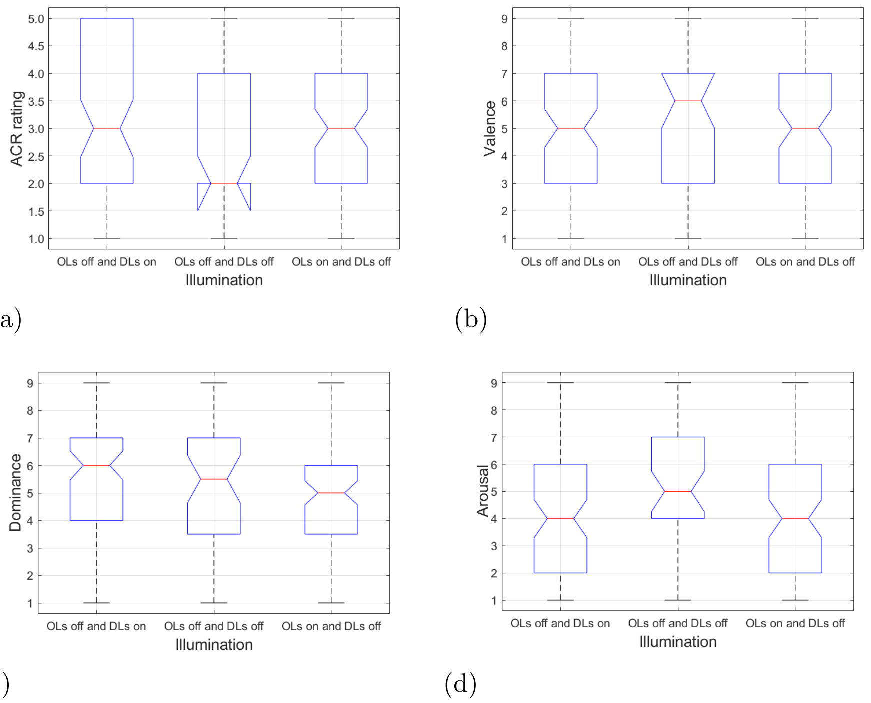


Figure 3.4: Mean Opinion Score (MOS) with 95% confidence interval (CI).

Multivariate Analysis of variance

The effects of ambient illumination and noise on ACR ratings and SAM scores (valence, dominance and arousal) are shown respectively in Fig. 3.5 and Fig. 3.6. Results demonstrate that the three different illumination conditions do not provide significant different effects on the four metrics. Indeed, the values assumed are within similar ranges for each illumination condition. In contrast, the effect of the noise is highly relevant on ACR rating and valence as the presence/absence of noise produces two separated groups of ACR ratings and valence scores. Specifically, the presence of noise greatly decreases the perceived QoE and valence. Instead, the noise does not seem to have a relevant impact on arousal and dominance.

In order to statistically investigate the effects of ambient illumination and noise on ACR ratings and SAM scores, I computed two-way repeated-measures MANOVA, which allows to identify differences between groups of collected scores based on the variation within and between different groups. Table 3.2 shows the effect of the two independent variables (illumination and noise) and of their combined effect on ACR rating, valence, dominance and arousal. It can be noticed that only the noise



(a) (b) (c) (d)

Figure 3.5: Effects of illumination on: (a) ACR-rating; (b) valence; (c) dominance; (d) arousal.

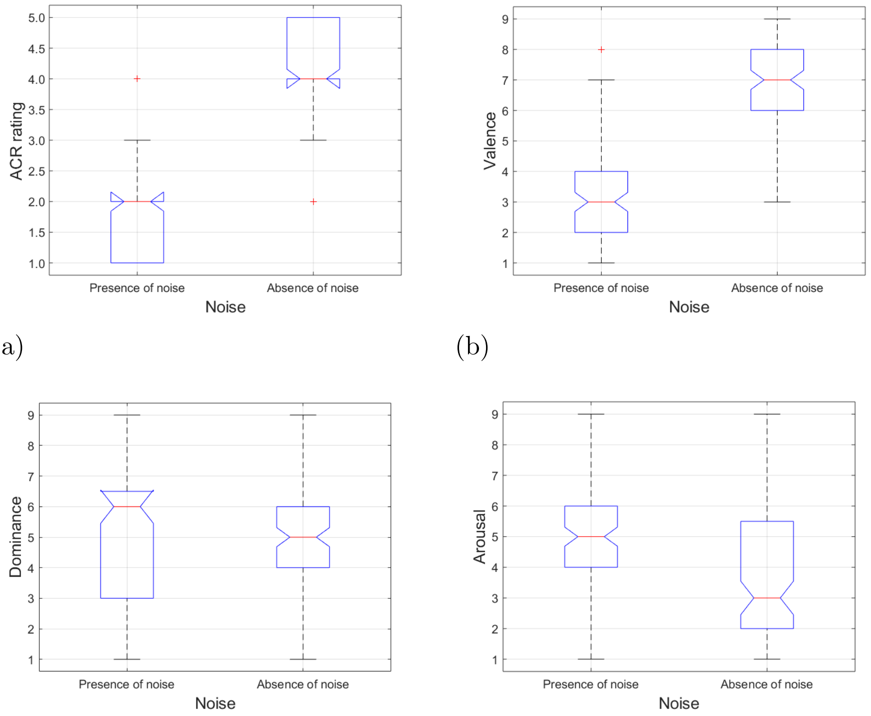
Metric		**Illumination**	**Noise**	**Illumination∗Noise**
ACR rating	F	1.75	494.85	0.18
	p	0.55	< 0.001	0.83
Valence	F	3.00	193.56	0.14
	p	0.05	< 0.001	0.87
Arousal	F	1.54	23.52	0.14
	p	0.22	< 0.001	0.71
Dominance	F	1.11	0.09	1.72
	p	0.33	0.77	0.19

Table 3.2: MANOVA analysis results.

shows a large effect on ACR rating ($F = 494.85$, $p < 0.001$), valence ($F = 193.56$, $p < 0.001$) and arousal ($F = 23.52$, $p < 0.001$). However, no effects of ambient illumination resulted on the four metrics. Same result is obtained for the combined effect of noise and illumination. Therefore, statistical results emphasise that only the ACR rating, valence and arousal groups related to the *absence of noise* condition are significantly different from those related to the *presence of noise* condition.

Furthermore, I computed the Pearson correlation between:

- ACR ratings and valence: 0.82;



(a)

(b)

(c)

(d)

Figure 3.6: Effects of noise on: (a) ACR-rating; (b) valence; (c) dominance; (d) arousal.

- ACR ratings and dominance: 0.02;
- ACR ratings and arousal: -0.36.

These results show that ACR ratings and valence scores are correlated; indeed, both of these metrics assume lower values in presence of noise whereas greater values are assumed when the noise is absent (see Fig. 3.6 (a)-(b)). Such a result highlights the importance of the emotions felt by the participant on the perceived quality during the watching/reading experience. By contrast, arousal and dominance are not correlated with ACR ratings.

QoE prediction

The subjective ratings provided by the participants have been divided into two datasets: a training dataset composed of the ratings provided by 14 participants (70% of the total) and a validation dataset composed of the ratings provided by the remaining 6 participants (30% of the total). The training/validation dataset selection followed a 5-fold cross-validation configuration. From the training dataset, I computed the MOS_T values, while from the validation dataset, I computed the MOS_V values.

A linear parametric model has been investigated to define a quality prediction model based on ambient illumination and noise stimuli. It is defined as follows:

$$MOS_p = a_n \cdot Noise + a_i \cdot Illumination + K \tag{3.1}$$

where $a_n = -2.341$, $a_i = 0.089$ and $K = 4.184$ are the coefficients obtained computing the linear regression between the stimuli and the MOS_T scores. Noise and illumination can assume the states described in Section 3.2.3. Finally, I computed the Pearson correlation between the MOS_v values and the MOS_p values predicted with the model in Eq. (3.1), which is 0.97. This result indicates that the proposed model can be used to predict the perceived QoE when the states of noise and illumination context stimuli are known.

3.2.6 Discussion

The effects of ambient illumination and noise on video watching and reading experiences have been investigated. The following findings are drawn:

- The presence of annoying noise has a relevant negative impact on user's QoE and valence;
- Ambient illumination has a slight impact on QoE and valence. Similarly to [NDKAK12] and [JAPM18], soft light effect is preferred than bright light ambient. However, the participants tended to ignore illumination conditions when the noise was present because this was too much annoying and predominant;
- QoE and valence are correlated. It seems that for this specific experiment the perceived quality is greatly influenced by the result of the emotions felt during the multimedia experience.

These results are interesting and open us to future challenges. The objective is to organize further experiments considering separately the effect of ambient illumination and noise to better identify their impact on QoE. Also, besides quality evaluation, I aim to implement control actions as a reply to the external stimuli. The room may be equipped with a noise sensor to detect annoying noise overlapping with the audio track of the watched video. This information may drive a smart room controller to implement some actions, such as increasing the volume of the audio track or suggesting the user to wear headphones. With regard to ambient illumination, people living in the same house may have different preferences about illumination conditions when watching videos. Controllable dome lights may be used to set an illumination being a trade-off among each person's preference. Furthermore, those people may be automatically recognized using a camera which records the TV viewers. Moreover, recording the face of the viewers may also be useful to predict the QoE from viewer emotions as results demonstrated the strong correlation between QoE and valence. If the viewer's face expresses a negative emotion, probably his perceived QoE is low. Emotion information may then be matched with context information sensed by IoT devices and used to drive QoE management. Machine learning may also be used to improve the accuracy of QoE models [TPDL18].

3.3 Quality of Experience Eye Gaze Analysis On HbbTV Smart Home Notification System

Nowadays we have a lot of devices always on and wired that could manage different systems of the house. Starting from this idea and the past work on a smart room, I set an experiment with an HbbTV able to show pop-ups about the incoming people at the door, incoming calls, notifications of the washing machine and status changing of a cooking monitor system. Furthermore, using the pieces of information given by the previous works about the prediction of the QoE with facial expressions and gaze direction characteristics, I tested the efficiency showed by the gaze direction features to understand the engagement level of the user with the pop-up windows.

Television has acquired a central role in the home environment being able to deliver multimedia content to the end users. Several studies presented different solutions to manage resources in order to broadcast multimedia contents in a real environment [JM, DAF$^+$18, SJR18, FMP12].

Integrated broadcast-broadband (IBB) systems allow broadcasters combining traditional with new type of contents (i.e., due to multifunctional applications) to the end-user [IR]. Recent studies have started to consider the HbbTV as the central hub to handle notifications regarding the smart home devices [SGCAM17]. Several European broadcasters started to take advantage of these new opportunities. For example, the British Broadcasting Corporation (BBC) carried out a new testing system in order to help manufacturers in testing performance of HbbTV devices to allow synchronization between TV applications and companion screen services with a program or channel that is being watched on the TV [BBC]. Moreover, the HbbTV Certification Group developed an HbbTV Operator Application (OpApp) to allow TV broadcasters:

- to control the user experience on Smart TVs and Set Top Boxs (STBs);
- to enable service-related TV operator support on consumer owned devices;
- to improve services maintaining the same level of functionalities over operator owned and consumer owned devices.

The research community is proposing new IBB systems to integrate heterogeneous applications and services provided by devices placed in a common environment, performing different operations, and able to communicate using different communication standards [FPC$^+$16, OOI$^+$, MBSV19]. The HbbTV integrates and shows into the reproduced multimedia TV contents the notifications coming from the various smart home devices, such as the smartphone, the video door-phone, the presence sensors, the washing machine, etc. In this way, the user does not have to control different devices while enjoying the preferred TV contents but the important notifications are directly shown on the TV screen [GPAF18].

Therefore, there is a need to evaluate the impact on-screen notifications would have on the user's perceived Quality of Experience (QoE) [LCMP12]. Indeed, in some cases notifications may disturb the user and negatively impact the QoE perceived for the watched video contents. For these reasons, in this work I present the results of a subjective quality assessment involving 30 people, aimed to evaluate the impact of on-screen notifications while they are watching TV video contents. Moreover, test participants were monitored while watching the video using the eye's gaze direction to measure the *attention* of the participants towards notifications.

Specifically, I considered notifications as pop-ups appearing on the TV screen while users are watching video contents. The impact of different pop-up characteristics, namely sound and content type, has been investigated. Three different options are considered for sound, i.e., no sound, notification and alarm, whereas the pop-up content mentions to the type of smart home devices sending the notification.

Besides asking users regarding the overall perceived QoE and usefulness/appreciation of the different pop-up characteristics, I aim to obtain interesting results from the analysis of user's gaze direction as this may reveal unconscious behaviour of the users.

3.3.1 Related Works

The majority of the QoE studies regarding multimedia services are focused on evaluating the effects of the impairments affecting network transmission, media quality and application performance [FHTG10a, PBC16, SES$^+$15]. However, other studies considered also the impact of the context on QoE. [MHSK$^+$18] focused on the potential of enhancing QoE management mechanisms by exploiting valuable context information whereas [LHM$^+$18] investigated the value of context-awareness in bandwidth-challenging HTTP adaptive streaming scenarios. [PFA19] focused on impairments provided by the surrounding environment, such as lights and noise, while watching a video on the TV. All these aforementioned studies investigated additional stimuli that are not linked to the multimedia content but that could impact on the overall QoE perceived by the end user.

Pop-up notifications may also have an impact on the perceived QoE of a multimedia content. The studies in [IH10, WMW15] investigated the reactions of the users to screen notifications. It resulted that notifications are disruptive for the subject focus. However, estimating a person's focus of attention is a challenging task. In [SYW01], authors have developed a system to predict focus of attention in a meeting situation from acoustic and visual information. An omnidirectional camera was employed to simultaneously track participants' faces around a meeting table and neural networks were used to estimate their head poses. In addition, microphones were used to detect who was speaking. In [RA14], an eye-tracking methodology was used to assess the gaze path of users in goal-directed and free-viewing tasks when viewing e-commerce pages with advertising banners. The objective was to identify the position where a banner becomes blind for the users. The study in [EPCZ10] has shown that distortions located in salient video regions have a significantly higher impact on quality perception as compared to distortions in non-salient video regions. For this reason, gaze direction is often integrated into image and video quality metrics with the objective to further improve their quality prediction performance [BBE$^+$09].

In this work, I extend the limited research regarding the impact of pop-up notifications on video quality of experience by considering different pop-up contents and sounds and by employing the gaze direction to assess the user's attention towards the appeared pop-up.

3.3.2 Methodology

The objective of this research is to investigate the effects of pop-up appearance on the user's QoE by considering both the user's subjective perception and the eye's gaze direction. From subjective perceptions I aim to collect insights regarding the annoyance and utility provided by pop-up appearance on the users based on the pop-up content type. Indeed, notifications from different smart home devices may have a different impact on user's perceptions. Some may be considered useful whereas other annoying. Moreover, 3 different options for the pop-up sounds have been considered, i.e., no sound, notification and alarm. No sound means that the pop-up just appears silently on the video without any sound (audio signal). The notification sound is a short sound similar to a message notification in a smartphone whereas the alarm sound is a short louder sound similar to a siren. Finally, to further evaluate user's perception, I measured and analyzed the eye's gaze direction to investigate the user's attention towards the pop-up. The aim is to understand whether different content types and positions may stimulate different levels of attention to the users.

Throughout the next sections, I shortly describe the methodology followed for tests presenting the HbbTV system, the test environment and setup, the subjective quality assessment, and the eye's gaze assessment.

HbbTV System

I developed a proof of concept based on the use of a smart home IoT gateway able to manage all the devices available in the home environment. The gateway is provided as a standalone Single Board Computer (SBC) device (i.e., Raspberry Pi) which operates in the same network as the TV set [JAPM, JAPM18]. The HbbTV terminal is a commercially available Nvidia Shield TV device [nvi] running Android TV 8.0 [and] on which a custom Live Channels application with HbbTV 2.0.1 support was deployed. Finally, a wireless Sensor/Actuator Networks (WSANs) to allow communication with Lo-RaWan, ZigBee, BLE, and WiFi sensors/actuators.

Test Environment and Setup

The test was performed at the QoE Lab of the Department of Electrical and Electronic engineering (DIEE) of the University of Cagliari, Italy. The QoE Lab is a 4 × 4 × 2.70 m (l × w × h) equivalent to a smart home living room equipped with a UHD 4K Flat Smart TV with a 60-inch diagonal and Wi-Fi Internet connection. Fig. 3.3 shows a picture of the room where the experiment was conducted. The tests involved 30 participants, who sat down on the sofa in front of the TV and singularly watched and rated the test video sequences enhanced with pop-up notifications. The participants were selected according to their specific preferred video content (i.e., action movie), which was the content type selected for all the considered test video sequences; however, all 12 video scenes were different. This choice is justified considering that the goal of the performed test disregards video characteristics.

For tests, five types of notification contents have been considered:

- **Washing machine**: the washing machine finishes/changes its program and communicates this information with a pop-up notification on the TV screen;
- **Video doorbell**: the doorbell system detects the presence of a guest and redirects the information with a video notification (obtained by the camera of the doorbell system) on the TV screen;
- **Cooking monitoring system**: the cooking monitoring system redirects information about critical/timing status changing of the cooking environment with a pop-up notification on the TV screen;

For each notification content type, three different options for the pop-up sound have been considered:

- **No sound**: the pop-up appears silently on the video without a sound;
- **Notification**: the notification sound is a short sound similar to a message notification in a smartphone;
- **Alarm**: the alarm sound is a short louder sound similar to a siren.

As an example, Figure 3.7a presents the HbbTV App displaying a video doorbell ring notification, while 3.7a.1 shows a zoom view of the displayed notification.

Finally, the notification display time was fixed to 5 seconds. This value was obtained by preliminary evaluation tests. Table 3.3 summarizes all the considered experiment conditions (ECs).

Subjective quality assessment

In total, 30 Italian participants, 10 females and 20 males between 20-50 years old, were recruited for the subjective quality assessment. All subjects reported normal or corrected-to-normal vision.

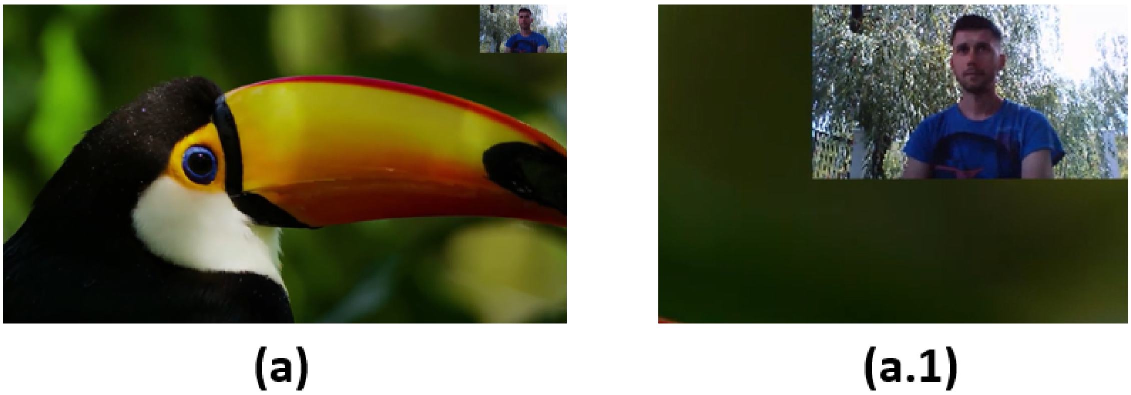


Figure 3.7: Notification display: video dorbell (a) and zoom view of the video dorbell (a.1)

EC	Video Genre	Pop-up Content	Pop-up Sound
1	Documentary	Washing Machine	No
2	TV Show	Cooking	Yes (Alarm)
3	Movie	Cooking	Yes (Alarm)
4	Sport	Washing Machine	Yes (Notification)
5	TV Show	Video doorbell	Yes (Notification)
6	Sport	Video doorbell	Yes (Notification)
7	Documentary	Cooking	No
8	Sport	Cooking	Yes (Alarm)
9	TV Show	Washing Machine	Yes (Notification)
10	Movie	Video doorbell	No
11	Movie	Washing Machine	Yes (Notification)
12	Documentary	Video doorbell	Yes (Notification)

Table 3.3: Experiment conditions (EC).

Before conducting the assessment, the participants were explained about all the different ECs they had to evaluate. They were informed that videos would be shown with different content and sound options regarding pop-up appearances and that they would be asked to evaluate the overall perceived QoE. To give participants an overview of the different stimuli involved in the experiment, 4 training videos were shown at first, which introduced all different pop-up appearance options. The participants were also trained about the utilization of the Absolute Category Rating (ACR) quality scale employed to capture variation in perceived QoE. The rating scale used in the assessment is based on Mean Opinion Score (MOS) as defined in the [ITU]. The ITU-T Rec. P.911 defines five-level rating scale: 5-Excellent, 4-Good, 3-Fair, 2-Poor, 1-Bad. Each participant in the assessment asked to give his/her quality rating for each video sequences.

An ad-hoc platform has been developed to show the video sequences and to collect the participants' QoE ratings. The platform has been developed using the Bootstrap open source toolkit for developing Web applications: HTML5 was used for the front-end side and JavaScript for the back-end side. The WebRTC JavaScript library has been used for video streaming. The videos were played in full screen mode on the TV screen. The platform guides the user through the video playing and rating phases with simple and intuitive buttons/indications. After watching each video the participants were asked to rate the overall perceived QoE including the impact of the appeared pop-up. After the submission of the perceived quality value, the next video was automatically shown. The overall time needed to complete the assessments was about 15 minutes. It is important to highlight that the ECs in Table 3.3 were presented in a different random order to each participant.

Eye's gaze assessment

During the subjective assessment, the test participants were recorded with a video camera placed on the top of the TV. The video recording was automatically started (ended) by the platform described in Section 3.3.2 when the video test to be watched started (ended). The recorded videos of the test participants were then processed with the OpenFace software, which allows to extract the gaze direction features from the faces of the participants [WBZ$^+$15, BMR15]. In particular, the eye gaze directions that reflect the eye movement of the person were extracted, i.e., $GazeAngle_x$, and $GazeAngle_y$. These features provides precise information regarding the horizontal and vertical movements of the eyes. If a person is looking left-right this will results in the change of $GazeAngle_x$ whereas if a person is looking up-down this will result in change of $GazeAngle_y$. If a person is looking straight ahead both of the angles will be close to 0. Therefore, before starting the test, I calibrated the video camera for each test participant so as to obtain a value close to 0 for both the gaze angles when the person was watching at the center of the screen. This was possible because OpenFace can show the gaze angle values from a video stream acquired in real-time. Once the camera was calibrated the test could start.

In the data analysis process, I used the collected gaze angles to measure the '*attention*' of the user towards the pop-up. In particular, I define the user's *attention* as the ratio between the time the user's eye gaze was directed towards the pop-up (within the time slot of the pop-up appearance) and the total time the pop-up appeared (5 seconds). As an example, if the participant's eye gaze was directed towards the pop-up for 1 second, I measured an attention of 20% (1 second out of 5 seconds). In Section 3.3.3, I analyze and comment the impact of pop-up content type and sound on the user's attention.

3.3.3 Experiment Results

In this section, a discussion about the experiment results is given. Section 3.3.3 focuses on the results of the subjective quality assessment in terms of the Mean Opinion Score (MOS). Then, the impact of pop-up appearance on ACR ratings and user's attention is evaluated in Section 3.3.3 by means of the Analysis of Variance (ANOVA).

Mean Opinion Score

In Fig. 3.8, I show the MOS scores with 95% confidence interval (CI) computed for each of the 12 ECs listed in Table 3.3. In the figure, I highlighted some of the most relevant results by grouping them based on common experiment's variable. First of all, it can be noticed that, except for EC10, all ECs including pop-up appearing with a sound achieved the lowest MOS. In particular, the Cooking pop-ups appearing with the alarm sound (EC2, EC3, EC8) achieved the three lowest MOS, ranging from 2.8 to 3, where 3 is perceived sufficient quality. Therefore, the alarm sound seems to be the most annoying for the users, even if the Cooking pop-up warns them about critical/timing status changing of the cooking environment. The Cooking pop-up that achieved higher MOS (the second highest MOS) is the one without sound (EC7), meaning that this pop-up is much more appreciated without the alarm sound. After the Cooking, the Video Doorbell is the least appreciated pop-up content (EC12, EC6, EC5, EC10), regardless of the pop-up sound. Indeed, whereas for EC12, EC6 and EC5 the pop-up appearance is accompanied by a short notification sound, for EC10 the pop-up appears with no sound. But all these ECs achieved almost the same MOS (between 3.03 and 3.13). The more appreciated (or less annoying) pop-up content is the Washing Machine. However, the Washing Machine pop-ups appeared accompanied by a notification sound (EC11, EC9, EC4) achieved lower MOS than that without sound (EC1). The notification group achieved MOS between 3.2 and 3.63, related to more than sufficient quality. The pop-up with no sound achieved the highest MOS for the conducted test, i.e., 3.73, very close to a good perceived quality. Finally, from the obtained MOS results, the video genre does not seem to have any particular influence on the perceived user's QoE.

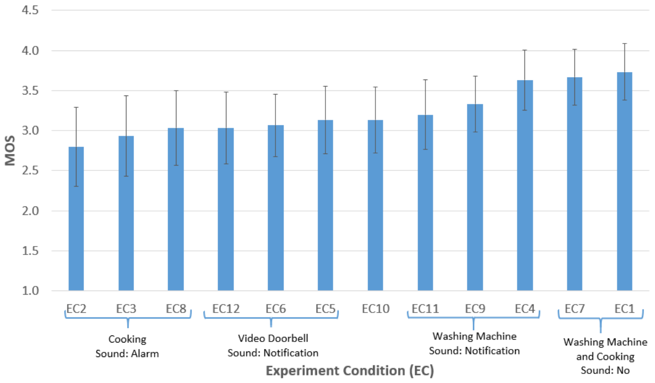


Figure 3.8: Mean Opinion Score (MOS) with 95% confidence interval (CI).

Summarizing the obtained MOS results, it can be seen that in general the achieved MOS is acceptable, ranging between 2.8 and 3.7. However, pop-ups accompanied by a sound were perceived as more annoying than those without a sound. This could be due by the fact that people appreciate in general the pop-up appearance because they can be informed about an event even while they are enjoying some multimedia content at the TV. However, they do not like this pop-up appearance to be too much intrusive with sounds that may disturb and decrease the perceived multimedia quality of the watched content. Finally, it is interesting to note that the confidence interval of the obtained MOS results is not negligible, meaning that pop-up perception is quite subjective among different people. Nevertheless, it must be noticed as the confidence interval decreases with the increases of MOS, which means that users agreed more on the good perceived quality provided by pop-ups without sound. On the other hand, pop-ups accompanied by sound are divisive among people.

Metric		**Pop-up Content**	**Pop-up Sound**
ACR rating	F	5.63	7.36
	p	0.04	< 0.001
User Attention	F	0.45	7.47
	p	0.635	< 0.001

Table 3.4: One-way ANOVA analysis results.

Analysis of variance

The ANOVA statistical test is typically used to determine whether data from several groups of a factor have a common mean. With regard to the conducted experiment, I employed the one-way ANOVA test to evaluate the effects of pop-up content and pop-up sound on ACR ratings and user's attention (as defined in Section 3.3.2). The MATLAB software to perform the ANOVA test on the collected data has been used, whose results are shown in Table 3.4. From these results, it can be seen that

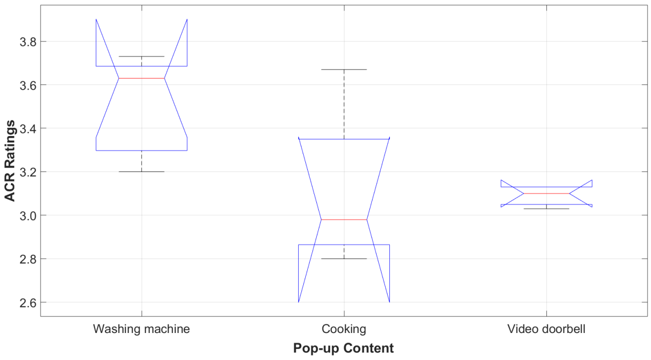


Figure 3.9: Effects of pop-up content on ACR ratings.

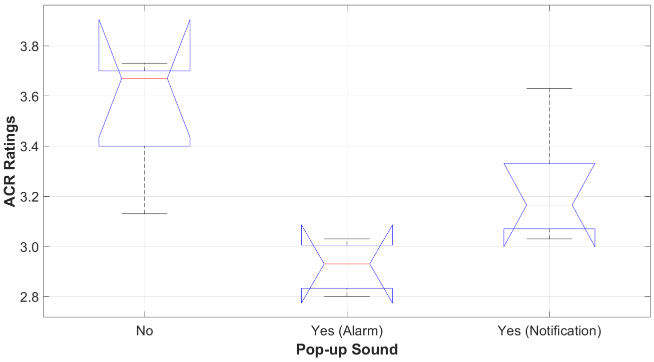


Figure 3.10: Effects of pop-up sound on ACR ratings.

only for the test regarding the effects of pop-up content on user's attention the null hypothesis is not rejected ($p = 0.635 > 0.05$) whereas for the other three tests the null hypothesis is rejected ($p < 0.05$). Therefore, it cannot be said that different pop-up contents are responsible for different levels of user's attention, but their impact is identical. This is also evident from Fig. 3.11.

With regard to the effects of pop-up content on ACR ratings (Fig. 3.9), the Washing Machine has means significantly different from Cooking and Video Doorbell. Indeed, it can be seen as, on average, the Washing Machine pop-up achieved higher ACR ratings than the other two pop-ups. Moreover, Cooking and Video Doorbell means are not significantly different and therefore their impact on ACR ratings is identical.

With regard to the effects of pop-up sound on ACR ratings (Fig. 3.10), the pop-up without sound has means significantly different from the pop-up with the Alarm sound. The pop-up with the Notification sound, instead, has not means significantly different from the other two groups. This means that only the extreme cases, i.e., No sound and Alarm sound, have an influence on ACR ratings (respectively a positive and a negative impact) whereas the Notification sound is on the middle ground.

Finally, with regard to the effects of pop-up sound on user's attention (Fig. 3.12),

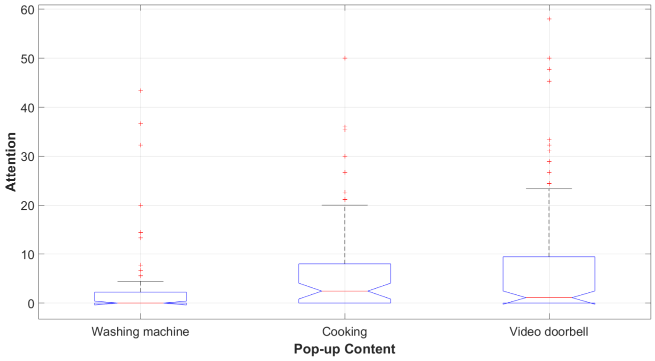


Figure 3.11: Effects of pop-up content on attention.

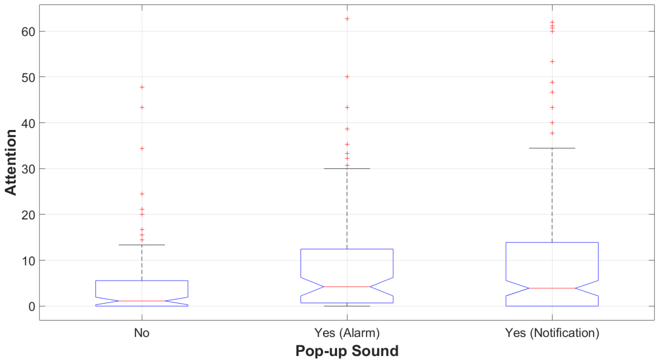


Figure 3.12: Effects of pop-up sound on attention.

the means of the pop-ups with sound (both Alarm and Notification) have means significantly different from the pop-up without sound. Therefore, user's attention is increased when pop-ups are accompanied by a sound. However, from these results, the impact on the two different sounds on user's attention is identical.

3.3.4 Conclusion

In this study, the impact of TV on-screen notifications on video quality of experience has been investigated. Notifications appeared in the form of pop-up containing different contents (Washing Machine, Cooking and Video Doorbell) and accompanied by different sounds (No sound, Notification, and Alarm). The results of the conducted subjective assessment have been analyzed both by computing the MOS and the ANOVA statistical test. The most relevant results are that in general pop-ups are appreciated (MOS ranging from 2.8 to 3.7) but pop-ups accompanied by a sound were perceived as more annoying achieving the lowest MOS. With regard to the pop-up content, the Washing Machine was the most appreciated whereas Cooking and Video Doorbell achieved lower MOS. These results were also confirmed by the ANOVA test. Moreover, it has been shown as different pop-up contents have identical impact on user's attention whereas different pop-up sounds have different impact on user's attention. In particular, attention is increased when pop-ups are accompanied by a sound.

Part IV

Conclusions and future works

Chapter 4

Conclusions and future works

In general, the thesis has been focused on the efficacy of FER adoption to evaluate the perceived QoE in a video-streaming session, starting from FER's DA rules to the definition of the framework ending with the adoption of the facial expressions in smart scenarios. In the following 3 paragraphs, a deeper explanation of each work is given.

In Part 1, a Convolutional Neural Network (CNN) able to classify the seven principal human emotions (neutral, happy, sad, angry, fear, disgust and surprise) has been used. Due to the hard task of recognize a categorized emotion from the face, different geometric data-augmentation (DA) techniques were studied to perform the emotion classification accuracy. The geometric DA techniques have the advantage to be fast and easy to be applied. In this study, to understand the efficiency of the geometric DA methods, the efficiency between DA methods and a Generative Adversarial Network (GAN) trained to create a new facial expression from scratch has been tested. However, it is not always easy to understand which DA technique may be more convenient for Facial Emotion Recognition (FER) systems because most of the state-of-the-art experiments use different settings, which makes the impact of DA techniques not comparable. To advance in this direction, I evaluated and compared the impact of using well-established DA techniques on the emotion recognition accuracy of a FER system based on the well-known VGG16 CNN neural network. In particular, both geometric transformations and a GAN to increase the number of training images have been considered. I have performed cross-database evaluations: training with the "augmented" KDEF DB and testing with two different DBs (CK+ and ExpW). The best results are obtained combining the horizontal reflection, the translation and the GAN, bringing to an accuracy increase of approximately 30%. This outperforms alternative existing approaches, even if other techniques could be comparable relying on larger DBs.

In Part 2 I investigated the effects of visual degradations on quality perception and emotional state of participants who were exposed to a series of short video clips. After each video playback, participants had to decide whether a certain event happened in the video. For data collection, subjective measures of quality and emotion were complemented by behavioural measures derived from capturing participants' spontaneous facial expressions. For data analysis, two general approaches were combined. First, a multivariate analysis of variance approach allowed to examine the effects of visual degradation factors on perceived quality and subjective emotional dimensions. It mainly revealed that perceived quality and emotional valence were both sensitive to degradation intensity, whereas the impact of degradation length was limited when task-relevant video content had already been obscured. Second, using a machine learning approach, an automatic Video Quality of Experience (VQoE) prediction system based on the recorded facial expressions was derived, demonstrating a strong correlation between facial expressions and perceived quality. Hereby, estimates of VQoE might be delivered in an objective, continuous and concealed manner, thus diminishing any further need for subjective self-reports. The analysis of recordings of participants' facial expressions during video playback enabled the construction of an automatic VQoE prediction system by following a machine learning approach. The obtained results show that emotional facial expressions are correlated with perceived quality, meaning that automatic, non-obtrusive VQoE prediction indeed remains a feasible goal. Since the data comes from two different datasets, obtained with different time length experiments, a metric to merge the dataset features has been developed. This metric allows to compare the facial features of video recordings of different length. Thus, by the training of different classifiers with the new bigger dataset, the study highlights that the best performance has been obtained with the k-NN, obtaining a prediction accuracy as high as 93.9 that outperforms the state-of-the-art achievements. As future works an open source software can be developed to help the QoE community to test this methodology. Other tests could also be done in various context, such as the VoIP calls or the Video calls.

In Part 3 other factors that could influence the QoE have been considered. in fact, visual impairments are not the only problems that can affect the perceived QoE during a multimedia session. Therefore, the environment has been studied, giving a definition of smart-city QoE modelling and introducing different experiments in a smart-room environment. As first step, I organized an experiment considering different combinations of ambient illumination and introducing a disturbing noise. Then, a subjective quality assessment has been conducted involving 20 people who were asked to rate the perceived QoE using the 5-level Absolute Category Rating (ACR) quality scale and to express their emotions completing the Self-Assessment Manikin (SAM) questionnaire. From this study has been discovered that: the presence of annoying noise has a relevant negative impact on user's QoE and valence; the soft light effect is preferred than bright light ambient, however, the participants tended to ignore illumination conditions when the noise was present because this was too much annoying and predominant; QoE and emotion valence are correlated. It seems that for this specific experiment the perceived quality is greatly influenced by the result of the emotions felt during the multimedia experience. The results validated our initial hypothesis.

Furthermore, a study with 30 participants to analyse the impact of TV on-screen notifications on video quality of experience has been investigated. Notifications appeared in the form of pop-up containing different contents (Washing Machine, Cooking and Video Doorbell) and accompanied by different sounds (No sound, Notification, and Alarm). In this test the gaze direction has been used to understand the level of engagement of the pop-up notification. The results of the conducted subjective assessment have been analysed both by computing the MOS and the ANOVA statistical test. It has been shown as the pop-up contents have identical impact on user's attention whereas different pop-up sounds have different impact on user's attention. In particular, attention is increased when pop-ups are accompanied by a sound. In future works, we aim to continue investigating how TV on-screen notifications impact the video QoE by also considering different positions and duration of notification appearance. Furthermore, additional sounds and contents may be considered so as to identify which of these may be appreciated by users or should be totally avoided while watching TV video contents.

For future work, it would be interesting to further investigate the impact of different types of content. Moreover, a deeper study concerning other human features, such as the voice's pitch could be considered in order to create a more efficient model.

Chapter 5

Appendix

Machine learning

ML today is a key component of information systems, capable of generalizing problems thanks to the inferring based on the recorded information extracted by the large databases. Though for its nature ML is a field of computer science, ML algorithms implementations differ from the typical computer science algorithms. From its definition, an algorithm is a series of steps that lead to a final result. In ML the approach is completely different. ML algorithm learns directly from different samples of the problem, and thanks to the statistical analysis application on inputs, it can output specific values that categorize the sample or that can describe the distance from this output.

The goal of ML is creating a model starting from input observations. These observations can be identified with three types of learning [AM14]:

- Supervised learning: the observations/samples are provided to the models in a tuple (f_i, t_i), in which f identifies the features set at the observation i and t means the target for the observation i described by its features.

- Unsupervised learning: the samples are presented only with descriptors. Since target elements are not presented, in this case the purpose is to find similarities between these values and to group similar data into clusters.

- Semi-supervised learning: the samples are presented as a small amount of labeled data (features with the correspondent target) and a large amount of unlabeled data during training(features without the correspondent target). It is a hybrid approach. The conjunction of unlabeled data and labeled one can considerably improve the learning accuracy because it makes the model more general.

ML models application depends on the type of problem that the researcher is facing. In QoE, ML methods use a set of features obtained by networks time-frame-stamps that are significant to the QoE. For example, considering a big set of observation of a network with a certain kind of problems that highlight a low QoE level could be helpful to train the ML model. This approach can extract the inference rules to automatically predict the low QoE level. To face this modelling problem it is mandatory choosing the best learning type. The learning type is not the only variable to take into account. In-fact in ML, there are a lot of models that perform only with a certain number of features because of their structure. Talking about supervised learning, that is the kind of ML approach used in this work, we can find different ML methods. Here there is a brief introduction to the most known methods:

- Linear Regression (LR): it is a linear approach used to find a correlation between the dependent variable and explanatory variables. It has an equation in the form $Y = X\beta + \epsilon$ where X is the explanatory variable and Y is the dependent variable, The slope of the line is β and ϵ is the value of y when $x = 0$ also called intercept.

- Multiple Linear Regression (MLR): it models the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. It is then very close to an LR model except for the number of dependent variables. Formally, it is described as follows: $Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + ... + \epsilon$.

- Decision Tree (DT): it is a categorical decision model based on the idea of a tree. Imaging to path a tree branch, at a certain point we will find a split between two different paths. Then we will decide for the right branch or the left one, based on what we need to find at the and of our path. The decision trees models work in the same way. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

- Support Vector Machine (SVM): it is a categorical or regression model. Focusing on the categorical classification, it tries to split the classes spreading the data space with hyperplanes. The idea is the same as a linear separation of the classes, but the usage of hyperplanes help the classification finding the best dimension to split the classes.

- K-Nearest Neighbors (K-NN):it is a type of instance-based learning. In this case, the model stores the position of the data, without creating a general model like the previous cases. The assignment of the class to each new sample is computed from a simple majority vote of the K nearest neighbors of each point. The more a sample is close to a cluster of points, the more is the probability of its belonging to their class.

Convolutional Neural Networks

CNN are a type of Deep Learning (DL)deep learning neural network. We can consider DL as a subset of ML set. It is also defined "hierarchical learning", in-fact the hierarchy of concepts allow machines to learn complicated concepts by building them out of simpler ones [Kim16]. The powerful of DL relies on its ability to learn from raw data such as pixels values of an image, and process them in a way that make possible to find a pattern within the image and to classify it. As explain for the ML general methods also the DL neural networks belong to the three categories of learning (supervised, unsupervised and semi-supervised) but because their usual multilayered stack architecture they need more data to be trained.

In this work, I will focus on CNN. As mention before, the CNN is a layered architecture oriented to the image classification that it is characterized by a convolutional layer. Convolution means that a kernel (or filter = small matrix) is applied to the input data to produce a feature map. In image processing the convolution is not a novel term, in fact, this kind of operation has been used for decades to detect edges in the image, to do the sharpening, the blurring and all the operations that require a kernel operation on the image. The convolution process in CNNs is used to extract the characteristics of the input images. This extraction highlights the components of the image that make it recognizable and classifiable [DSC20]. A simple CNN is structured as a sequences of steps:

- first step: it is composed of the input image layer. Technically it is a tensor with a $n \times (h \times w \times d)$ shape where n is the number of images, h w and d are respectively the height, the weight and the depth of the images.

- second step: it is composed of the convolutional layers and the pooling layers. Images in a convolutional layer are organized in feature maps with $n \times (h \times w \times d)$ notation. Each image (now called unit) is connected to local patches in the feature maps of the previous layer through a set of weights called a filter bank. The result of this local weighted sum is then passed through a non-linearity such as a Rectified Linear Unit (ReLU). The ReLU effectively changes the negative values setting them to zero advantaging the increasing of nonlinear porperties of the decision function.

- third step: applying osf the fully-connected layer. After the series of convolutional and max-pooling layers, a fully connected layer link each neuron with the activations extracted from the previous layer.

- fourth layer: applying of Softmax layer. In the end, we find the loss layer. Typically the Softmax layer is used to predict a single class of N mutually exclusive classes. This layer specifies the error between the predicted class and the true label.

Chapter 6

List of publications related to the Thesis

- S. Porcu, A. Floris, M. Anedda, V. Popescu, M. Fadda and L. Atzori, "Quality of Experience Eye Gaze Analysis On HbbTV Smart Home Notification System," in IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, Paris, France, 2020.

- S. Porcu, A. Floris, J. Voigt-Antons, L. Atzori and S. Möller,"Estimation of the Quality of Experience during Video Streaming from Facial Expression and Gaze Direction," in IEEE Transactions on Network and Service Management.

- S. Porcu, A. Floris and L. Atzori, "Towards the Evaluation of the Effects of Ambient Illumination and Noise on Quality of Experience," 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX),Berlin, Germany, 2019, pp. 1-6.

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

- S. Porcu, A. Floris and L. Atzori, "Towards the Prediction of the Quality of Experience from Facial Expression and Gaze Direction," 2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), Paris, France, 2019, pp. 82-87.

- A. Floris, S. Porcu and L. Atzori, "Quality of Experience Management of Smart City services," 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX), Cagliari, 2018, pp. 1-3.

List of Figures

List of Tables

Bibliography

[AA20] 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*, 17(1):361–374, 2020.

[AASM16] S. Arndt, J. N. Antons, R. Schleicher, and S. M¨ller. Using electroencephalography to analyze sleepiness due to low-quality audiovisual stimuli. *Signal Processing: Image Communication*, 42:120–129, March 2016.

[ABSM18] 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)*, pages 1–6, 2018.

[AD14] N. Aifanti and A. Delopoulos. Linear subspaces for facial expression recognition. *Signal Processing: Image Communication*, 29(1):177 – 188, 2014.

[AM14] S. Aroussi and A. Mellouk. Survey on machine learning-based qoe-qos correlation models. In *2014 International Conference on Computing, Management and Telecommunications (ComManTel)*, pages 200–204, 2014.

[and] Android live tv.

[ARM14] S. Arndt, JN. Radun, J. 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*, pages 281–285. IEEE, 2014.

[ASA+12] J. N. Antons, R. Schleicher, S. Arndt, S. Möller, A. Porbadnigk, and G. Curio. Analyzing speech quality perception using electroencephalography. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):721–731, 2012.

[AWZ12] B. Abhishek, W. Wanmin, and Y. Zhenyu. Quality of experience evaluation of voice communication: an affect-based approach. *Human-centric Computing and Information Sciences*, 2(1), 2012.

[Bao20] Do Quoc Bao. Image Blur Metric, MATLAB Central File Exchange, 2020.

[BBC] BBC. Release of HbbTV / DVB companion synchronisation tools and streams - BBC r&d.

[BBE$^+$09] 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*, 3(2):266–279, 2009.

[BDF16] B. Barabino and M. Di Francesco. Characterizing, measuring, and managing transit service quality. *Journal of Advanced Transportation*, 50(5):818–840, 2016.

[BJGM19] 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*, 7:74511–74527, 2019.

[BL94] M. M. Bradley and P. J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994.

[BM15] Ó. Ballesteros, L. Álvarez and J. Markendahl. Quality of Experience (QoE) in the smart cities context: An Initial Analysis. In *IEEE First Int. Smart Cities Conference (ISC2), 2015*. IEEE, 2015.

[BMR15] T. Baltruaitis, 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)*, volume 06, pages 1–6, 2015.

[BZLM18] 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)*, pages 59–66, 2018.

[CDF$^+$14] 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*, 11(4):441–457, Dec 2014.

[CDLN07] 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*, pages EI 6492–16, 2007.

[CDW$^+$17] 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)*, pages 1–6, 2017.

[CHW$^+$19] Q. Chu, M. Hu, X. Wang, Y. Gu, and T. Chen. Facial Expression Recognition Based on Contextual Generative Adversarial Network. In *2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pages 120–125, 2019.

[Cis20] Cisco. Cisco visual networking index: Forecast and trends, 2017–2022 white paper, 2020.

[CM20] E. Cipressi and M. L. Merani. An Effective Machine Learning (ML) Approach to Quality Assessment of Voice over IP (VoIP) Calls. *IEEE Networking Letters*, pages 1–1, 2020.

[CPTP15] 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*, 15(6):1443–1456, 2015.

[CSW$^+$16] 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*, 13(2):181–196, June 2016.

[DAF$^+$18] A. Domínguez, M. Agirre, J. Flörez, A. Lafuente, I. Tamayo, and M. Zorilla. Deployment of a hybrid broadcast-internet multi-device service for a live tv programme. In *IEEE Transactions on Broadcasting*, volume 64, pages 153–163, 2018.

[DSC20] S. Do, K. D. Song, and J. W. Chung. Basics of deep learning: A radiologist's guide to understanding published radiology articles on deep learning. *kjr*, 21(1):33–41, 2020.

[dSP15] F. A. M. da Silva and H. Pedrini. Effects of cultural characteristics on building an emotion classifier through facial expression analysis. *Journal of Electronic Imaging*, 24(2):1 – 9 – 9, 2015.

[DT15] C. Ding and D. Tao. Robust Face Recognition via Multimodal Deep Face Representation. *IEEE Trans. on Multimedia*, 17(11):2049–2058, Nov 2015.

[EDM$^+$17] 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*, 11(1):6–21, 2017.

[EF71] P. Ekman and W. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124–129, 1971.

[EF78] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* San Francisco: Consulting Psychologists Press, 1978.

[EPCZ10] U. Engelke, R Pepion, P Le Callet, and H. Zepernick. Linking distortion perception and visual saliency in H.264/AVC coded video containing packet loss. In *Proc. SPIE*, volume 7744, 2010.

[FHTG10a] M. Fiedler, T. Hoßfeld, and P. Tran-Gia. A generic quantitative relationship between quality of experience and quality of service. *IEEE Network*, 24(2):36–41, 2010.

[FHTG10b] M. Fiedler, T. Hoßfeld, and P. Tran-Gia. A generic quantitative relationship between Quality of Experience and Quality of Service. *IEEE Network*, 24(2):36–41, 2010.

[FMP12] M. Fadda, M. Murroni, and V. Popescu. An unlicensed indoor hdtv multi-vision system in the dtt bands. *IEEE Transactions on Broadcasting*, 58(3):338–346, 2012.

[FPC$^+$16] P. A. Fam, S. Paquelet, M. Crussière, J. Hélard, and P. Brétillon. Analytical derivation and optimization of a hybrid unicast-broadcast network for linear services. In *IEEE Transactions on Broadcasting*, volume 62, pages 890–902, 2016.

[GAC16] M. Amjad Iqbal G. Ali and TS. Choi. Boosted NNE collections for multicultural facial expression recognition. *Pattern Recognition*, 55:14 – 27, 2016.

[GBY$^+$14] 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)*, pages 989–993, 2014.

[GBY$^+$16] D. Ghadiyaram, A. C. Bovik, H. Yeganeh, R. Kordasiewicz, and M. Gallant. LIVE Mobile Stall Video database, 2016.

[GDRLV08] E. Goeleven, R. De Raedt, L. Leyman, and B. Verschuere. The karolinska directed emotional faces: A validation study. 22:1094–1118, 2008.

[GPAF18] C. Gavrilă, V. Popescu, M. Alexandru, and M. Fadda. Unifying the smart home experience through hbbtv-enabled devices. In *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting 2019, June 5th-7th, Jeju, South Korea*, 2018.

[GPAM$^+$14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[GXV$^+$12] W. Gu, C. Xiang, Y. V. Venkatesh, D. Huang, and H. Lin. Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. *Pattern Recognition*, 45(1):80 – 91, 2012.

[HBGL08] 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)*, pages 1322–1328, June 2008.

[HHM$^+$15] M. Hirth, T. Hoßfeld, M. Mellia, C. Schwartz, and F. Lehrieder. Crowdsourced network measurements: Benefits and best practices. *Computer Networks*, 90:85 – 98, 2015.

[HHVSK18] T. Hoßfeld, P. E. Heegaard, M. Varela, and L. Skorin-Kapov. Confidence Interval Estimators for MOS Values, 2018.

[HIRR09] H. F. A. Hasan, A. Ilias, R.A. Rahman, and M. Z. A. Razak. Service Quality and Student Satisfaction: A Case Study at Private Higher Education Institutions. *International Business Research*, 1(3):163–175, 2009.

[HKH$^+$14] 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*, 16(2):541–558, 2014.

[HM17] B. Hasani and M. H. Mahoor. Spatio-Temporal Facial Expression Recognition Using Convolutional Neural Networks and Conditional Random Fields. *CoRR*, abs/1703.06995, 2017.

[HSH$^+$11] 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*, pages 494–499, 2011.

[HYGS08] 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)*, pages 1322–1328, 2008.

[IH10] S. T. Iqbal and E. Horvitz. Notifications and awareness: A field study of alert usage and preferences. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 27–30, New York, NY, USA, 2010. ACM.

[IR] ITU-R. Integrated broadcast-broadband systems". in:report itu-r bt.2267-10.

[ITU] ITU-R BT.500-14: Methodologies for the subjective assessment of the quality of television images.

[ITU08] Subjective video quality assessment methods for multimedia applications. Recommendation ITU-T P.910, 2008.

[ITU12] General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays. Recommendation ITU-R BT.2022, 2012.

[ITU17] ITU. Models and tools for quality assessment of streamed media. Recommendation ITU-T P.1203, 2017.

[JAPM] L. Jalal, M. Anedda, V. Popescu, and M. Murroni. Qoe assessment for broadcasting multi sensorial media in smart home scenario. In *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), 6-8 June 2018, Valencia, Spain.*

[JAPM18] L. Jalal, M. Anedda, V. Popescu, and M. Murroni. Qoe assessment for iot-based multi sensorial media broadcasting. *IEEE Transactions on Broadcasting*, 64(2):552–560, 2018.

[JM] L. Jalal and M. Murroni. The impact of multi-sensorial media in smart home scenario on user experience and emotions. In *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting 2019, June 5th-7th, Jeju, South Korea.*

[JNBJ08] I. B. Mauss J. J. Gross M. E. Jabon C. A. C. Hutcherson C. Nass J. N. Bailenson, E. D. Pontikakis and O. John. Real-time classification of evoked emotions using facial feature tracking and physiological responses. *Int. Journal of Human-Computer Studies*, 66(5):303 – 317, 2008.

[KDSS19] 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*, 16(3):1113–1126, Sep. 2019.

[KHL$^+$14] 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)*, pages 2135–2139, 2014.

[Kim16] K. G. Kim. Book review: Deep learning. *Healthc Inform Res*, 22(4):351–354, 2016.

[KKKL18] H. Kim, Y. Kim, S. J. Kim, and I. Lee. Building Emotional Machines: Recognizing Image Emotions Through Deep Neural Networks. *IEEE Trans. on Multimedia*, 20(11):2980–2992, Nov 2018.

[LC12] K. U. Rehman Laghari and K. Connelly. Toward total quality of experience: A QoE model in a communication ecosystem. *IEEE Communications Magazine*, 50(4):58–65, 2012.

[LCK$^+$10] 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*, pages 94–101, 2010.

[LCL$^+$18] 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*, 56(9):110–117, 2018.

[LCMP12] P. Le Callet, S. Möller, and A. Perkis. Qualinet White Paper on Definitions of Quality of Experience (2012), 2012. European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, Version 1.2, March 2013.

[LD18] S. Li and W. Deng. Deep Facial Expression Recognition: A Survey. *CoRR*, abs/1804.08348, 2018.

[LdASOS17] A. Teixeira Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos. Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. *Pattern Recognition*, 61:610 – 628, 2017.

[Lee17] D. Lee. HEALTHQUAL: a multi-item scale for assessing healthcare service quality. *Service Business*, 11(3):491–516, 2017.

[LHM+18] E. Liotou, T. Hoßfeld, C. Moldovan, F. Metzger, D. Tsolkas, and N. Passas. *The Value of Context-Awareness in Bandwidth-Challenging HTTP Adaptive Streaming Scenarios*, pages 128–150. Springer, 2018.

[LHZL18] F. Lin, R. Hong, W. Zhou, and H. Li. Facial Expression Recognition with Data Augmentation and Compact Feature Learning. In *2018 25th IEEE Int. Conf. on Image Processing (ICIP)*, pages 1957–1961, 2018.

[LKB+19] 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*, 21(10):2589–2602, Oct 2019.

[LMR+17] K. Lekdioui, R. Messoussi, Y. Ruichek, Y. Chaabi, and R. Touahni. Facial decomposition for expression recognition using texture/shape descriptors and SVM classifier. *Signal Processing: Image Communication*, 58:300 – 312, 2017.

[LMSS20] 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)*, pages 152–159, 2020.

[MAAWI+19] F. Makhmudkhujaev, M. Abdullah-Al-Wadud, M. T. B. Iqbal, B. Ryu, and O. Chae. Facial expression recognition with local prominent directional pattern. *Signal Processing: Image Communication*, 74:1 – 12, 2019.

[MBL+15] 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*, 66:100 – 108, 2015. Pattern Recognition in Human Computer Interaction.

[MBSV19] D. Marfil, F. Boronat, A. Sapena, and A. Vidal. Synchronization mechanisms for multi-user and multi-device hybrid broadcast and broadband distributed scenarios. In *IEEE Access*, volume 7, pages 605–624, 2019.

[MCM16] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE*

Winter Conference on Applications of Computer Vision (WACV), pages 1–10, March 2016.

[MH19] S. Mustafa and A. Hameed. Perceptual quality assessment of video using machine learning algorithm. *Signal, Image and Video Processing*, 13(8):1495–1502, 2019.

[MHSK⁺18] F. Metzger, T. Hoßfeld, L. Skorin-Kapov, Y. Haddad, E. Liotou, P. Pocta, H. Melvin, V. A. Siris, A. Zgank, and M. Jarschel. *Context Monitoring for Improved System Performance and QoE*, pages 23–48. Springer, 2018.

[MR14] S. Möller and A. Raake. *Quality of Experience.* Springer, 2014.

[MZ5] K. Mitra, A. Zaslavsky, and C. Åhlund. Context-Aware QoE Modelling, Measurement, and Prediction in Mobile Computing Systems. *IEEE Transactions on Mobile Computing*, 14(5):920–936, 2015.

[MJ19] 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)*, pages 1–6, 2019.

[NDKAK12] G. Nur, S. Dogan, H. Kodikara Arachchi, and A. M. Kondoz. Assessing the Effects of Ambient Illumination Change in Usage Environment on 3D Video Perception for User Centric Media Access and Consumption. In Federico Alvarez and Cristina Costa, editors, *User Centric Media*, pages 60–68. Springer Berlin Heidelberg, 2012.

[nvi] Nvidia shield.

[OOI⁺] H. Ogawa, H. Ohmata, M. Ikeo, A. Fujii, and H. Fujisawa. System architecture for content-oriented iot services. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), 11-15 March 2019, Kyoto, Japan.*

[PBC16] P. Paudyal, F. Battisti, and M. Carli. Impact of video content and transmission impairments on quality of experience. *Multimedia Tools and Applications*, 75(23):16461–16485, Dec 2016.

[PFA19] S. Porcu, A. Floris, and L. Atzori. Towards the evaluation of the effects of ambient illumination and noise on quality of experience. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, June 2019.

[PGC$^+$17] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[PWBL17] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana. Enhancing CNN with Preprocessing Stage in Automatic Emotion Recognition. *Procedia Computer Science*, 116:523 – 529, 2017.

[PZB88] A. Parasuraman, Valarie A. Zeithaml, and Leonard L. Berry. SERVQUAL: a multiple item scale for measuring consumer perception of service quality. *Journal of Retailing*, 64(1):12–37, 1988.

[RA14] M. Resnick and W. Albert. The impact of advertising location and user task on the emergence of banner ad blindness: An eye-tracking study. *International Journal of Human–Computer Interaction*, 30(3):206–219, 2014.

[RBLC16] Y. Rai, M. Barkowsky, and P. Le Callet. Role of spatio-temporal distortions in the visual periphery in disrupting natural attention deployment. *Electronic Imaging*, 2016(16):1–6, 2016.

[RESD10] P. Reichl, S. Egger, R. Schatz, and A. D'Alconzo. The Logarithmic Nature of QoE and the Role of the Weber-Fechner Law in QoE Assessment. In *Proc. of the IEEE International Conference on Communications (ICC), 2010*. IEEE, 2010.

[RGR$^+$17] 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.

[RGR$^+$18] 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.

[RLC17] Y. Rai and P. Le Callet. Do gaze disruptions indicate the perceived quality of non-uniformly coded natural scenes? . In *Human Vision and Electronic Imaging 2017*, pages 104–109, 2017.

[SA14] R. Schleicher and J. N. Antons. Evoking Emotions and Evaluating Emotional Impact. In Sebastian Möller and Alexander Raake, editors, *Quality of Experience: Advanced Concepts, Applications and Methods*, pages 121–132. Springer International Publishing, Cham, 2014.

[SCW+19] 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)*, pages 688–695, 2019.

[SES+15] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia. A Survey on Quality of Experience of HTTP Adaptive Streaming. *IEEE Comm. Surveys Tutorials*, 17(1):469–492, 2015.

[SGCAM17] H. Sánchez, C. González-Contreras, J. E. Agudo, and M. Macías. IoT and iTV for Interconnection, Monitoring, and Automation of Common Areas of Residents. *Applied Sciences*, 7(7), 2017.

[SGM09] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803 – 816, 2009.

[SJR18] R. Sotelo, J. Joskowicz, and N. Rondán. An integrated broadcast-broadband system that merges ISDB-t with HbbTV 2.0. In *IEEE Transactions on Broadcasting*, volume 64, pages 709–720, 2018.

[SK19] C. Shorten and T. M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60, Jul 2019.

[SKVHC18] 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.*, 14(2s), May 2018.

[SLU+15] S. Morteza Safdarnejad, Xiaoming Liu, Lalita Udpa, Brooks Andrus, John Wood, and Dean Craven. Sports Videos in the Wild (SVW): A Video Dataset for Sports Analysis. In *Proc. of the IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2015.

[SNR15] M. Shafiuzzaman, N. Nahar, and M. R. Rahman. A proactive approach for context-aware self-adaptive mobile applications to ensure Quality of Service. In *2015 18th International Conference on Computer and Information Technology (ICCIT)*, pages 544–549, 2015.

[SSP03] P. Y. Simard, D. Steinkraus, and J. C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *Seventh Int. Conf. on Document Analysis and Recognition, 2003. Proceedings.*, pages 958–963, 2003.

[SYW01] R. Stiefelhagen, J. Yang, and A. Waibel. Estimating focus of attention based on gaze and sound. In *Proceedings of the 2001 Workshop on Perceptive User Interfaces*, PUI '01, pages 1–9, New York, NY, USA, 2001. ACM.

[SZ14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.

[TDL+15] 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*, 29(6):14–19, 2015.

[TLPM17] D. Tsolkas, E. Liotou, N. Passas, and L. Merakos. A survey on parametric QoE estimation for popular services. *Journal of Network and Computer Applications*, 77(1):1–17, 2017.

[TPDL18] 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*, 64(2):432–445, June 2018.

[TVMF+17] 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*, 24(6):736–740, 2017.

[TVMSL17] M. Torres Vega, D. C. Mocanu, S. Stavrou, and A. Liotta. Predictive no-reference assessment of video quality. *Signal Processing: Image Communication*, 52:20–32, 2017.

[UMMVA19] 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.

[VKA+11] S. Velusamy, H. Kannan, B. Anand, A. Sharma, and B. Navathe. A method to infer emotions from facial action units. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2028–2031, 2011.

[WBZ+15] E. Wood, T. Baltruaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling. Rendering of Eyes for Eye-Shape Registration and Gaze Estimation. In *IEEE ICCV*, pages 3756–3764, 2015.

[WHM11] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534, June 2011.

[WMW15] T. Westermann, S. Möller, and I. Wechsung. Assessing the relationship between technical affinity, stress and notifications on smartphones. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, MobileHCI '15, pages 652–659, New York, NY, USA, 2015. ACM.

[WSSE15] S. Wilk, S. Schönherr, D. Stohr, and W. Effelsberg. EnvDASH: An Environment-Aware Dynamic Adaptive Streaming over HTTP System. In *Proc. of the ACM Int. Conf. on Interactive Experiences for TV and Online Video*, TVX '15, pages 113–118. ACM, 2015.

[XH19] S. Xie and H. Hu. Facial Expression Recognition Using Hierarchical Features With Deep Comprehensive Multipatches Aggregation Convolutional Neural Networks. *IEEE Trans. on Multimedia*, 21(1):211–220, Jan 2019.

[Yar14] E. K. Yarimoglu. A Review on Dimensions of Service Quality Models. *Journal of Marketing Management*, 2(2):79–93, 2014.

[YLG$^+$18] 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*, 15(1):444–457, March 2018.

[YSH18] W. Yi, Y. Sun, and S. He. Data Augmentation Using Conditional GANs for Facial Emotion Recognition. In *2018 Progress in Electromagnetics Research Symposium (PIERS-Toyama)*, pages 710–714, Aug 2018.

[ZBO17] M. V. Zavarez, R. F. Berriel, and T. Oliveira-Santos. Cross-Database Facial Expression Recognition Based on Fine-Tuned Deep Convolutional Network. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 405–412, Oct 2017.

[ZLL$^+$18] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin. Emotion Classification with Data Augmentation Using Generative Adversarial Networks. In Dinh Phung, Vincent S. Tseng, Geoffrey I. Webb, Bao Ho, Mohadeseh Ganji, and Lida Rashidi, editors, *Advances in Knowledge Discovery and Data Mining*, pages 349–360, Cham, 2018. Springer International Publishing.

[ZLLT16] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. From facial expression recognition to interpersonal relation prediction. In *arXiv:1609.06426v2*, 2016.

[ZSX$^+$19] Y. Zhang, B. Sun, Y. Xiao, R. Xiao, and Y. Wei. Feature augmentation for imbalanced classification with conditional mixture WGANs. *Signal Processing: Image Communication*, 75:89 – 99, 2019.

[ZZC$^+$16] 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*, 18(12):2528–2536, Dec 2016.

[ZZL$^+$17] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Li. Sfd: Single shot scale-invariant face detector. 10 2017.

La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005),
Fondo Sociale Europeo, Azione I.1 “Dottorati Innovativi con caratterizzazione Industriale”

UNIONE EUROPEA
Fondo Sociale Europeo

Ministero dell'Università
e della Ricerca

PON
RICERCA
E INNOVAZIONE
2014 - 2020