Towards the Prediction of the Quality of Experience from Facial Expression and Gaze Direction

Simone Porcu, Alessandro Floris, and Luigi Atzori
DIEE, University of Cagliari, UdR CNIT of Cagliari, Italy
simone.porcu@diee.unica.it, alessandro.floris@diee.unica.it, l.atzori@ieee.org

***Abstract*—In this paper we investigate on the potentials to implicitly estimate the Quality of Experience (QoE) of a user of video streaming services by acquiring a video of her face and monitoring her facial expression and gaze direction. To this, we conducted a crowdsourcing test in which participants were asked to watch and rate the quality when watching 20 videos subject to different impairments, while their face was recorded with their PC's webcam. The following features were then considered: the Action Units (AU) that represent the facial expression, and the position of the eyes' pupil. These features were then used, together with the respective QoE values provided by the participants, to train three machine learning classifiers, namely, Support Vector Machine with quadratic kernel, RUSBoost trees and bagged trees. We considered two prediction models: only the AU features are considered or together with the position of the eyes' pupils. The RUSBoost trees achieved the best results in terms of accuracy, sensitivity and area under the curve scores. In particular, when all the features were considered, the achieved accuracy is of 44.7%, 59.4% and 75.3% when using the 5-level, 3-level and 2-level quality scales, respectively. Whereas these results are not satisfactory yet, these represent a promising basis.**

***Index Terms*—Quality of Experience, Facial expression, Gaze direction, Crowdsourcing, Machine learning.**

I. Introduction

Quality of Experience (QoE) evaluation is becoming more and more important for the successful deployment of multimedia services, due to a constantly increase in the users expectations. The collection of user's subjective perceived quality and feedback is of paramount importance to identify root causes of quality degradation and take necessary corrective actions in service management. The user QoE for multimedia services is obtained by asking people to provide a measure of their perceived quality. However, 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.

For these reasons, alternative approaches have been studied in the last years, which are mostly based on psychophysiology measures (e.g., electroencephalography (EEG), gaze direction) to identify implicit rather than explicit responses to physical stimuli (user's cognitive state) and thus to overcome the problem of potentially misleading rating scales and conscious decision making [1]. The objective of these approaches is not to replace subjective tests but to support them to provide deeper insights into high-level QoE attributes. As a further step taken towards this direction, recent studies (e.g., [2], [3]) have investigated the correlation between human facial expressions and the viewer's sentimental state when watching a video sequence. Indeed, recent findings in neuroscience suggest that a human's sentimental state can be conveyed by facial expressions and body movements [4].

The subject of this paper falls into this area, with a specific focus on the prediction of the QoE based on human facial expressions and gaze direction. The objective is to investigate 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 to bother the subjects with questions to collect opinions and feedback. We specifically focused on the scenario of perceived quality estimation during video streaming sessions. To this, we conducted a crowdsourcing test in which participants were asked to watch and rate the quality of 20 videos subject to impairments caused by long initial delays and re-buffering events. The reason we relied on crowdsourcing is that it allows for the involvement of a large number of subjects in a shorter time with respect to lab experiments, although it has the intrinsic drawback introduced by performing the test in a non-controlled environment [5]. The test was conducted using the crowdsourcing platform Amazon Mechanical Turk (MTurk)[1]. The test participants were asked to watch each of the 20 videos while their face was recorded with their PC's webcam.

The Action Units (AU) features, which represent the contractions of specific facial muscles, were considered as the features to express the facial expression of the viewer. The position of the eyes' pupil was the feature considered to evaluate the gaze direction of the viewer. Those features were then used, together with the respective QoE values provided by the participants, to train three machine learning (ML) classifiers, namely, Support Vector Machine (SVM) with quadratic kernel, RUSBoost trees and bagged trees, with the aim to predict the perceived QoE on the basis of the viewer's face expression and gaze direction. Specifically, we considered two prediction models: i) *AUtoQoE*: takes as input the AU

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

[1] https://www.mturk.com/

features; ii) *AU&GAZEtoQoE*: takes as input the AU features and the position of the eyes' pupil.

We validated the QoE prediction models using three different quality scales: 5-level, 3-level and 2-level scales. The performance metrics used to compare the considered classifiers using the 3 considered quality scales are the accuracy, the sensitivity and the area under the curve (AUC) scores. The RUSBoost trees achieved the best results in terms of all the performance metrics. In particular, the achieved accuracy is respectively 44.7%, 59.4% and 75.3% for the AU&GAZEtoQoE model using the 5-level, 3-level and 2-level quality scales.

The paper is structured as follows. Section II discusses the major related works in this area. In Section III an overview of the proposed system is presented. Section IV presents the methodology followed by the proposed study whereas Section V shows the results achieved by the proposed QoE prediction models. Finally, Section VI concludes the paper.

II. Past works

Psychophysiology is concerned with the physiological bases of perceptual and cognitive processes. As such, psychophysiological methods measure implicit responses to physical stimuli and thus overcome the problem of potentially misleading rating scales and conscious decision making. The drawback is that these methods are obtrusive to users and need special equipment or devices [1]. Particularly relevant to QoE-based neurophysiological investigations is electroencephalography (EEG), which is a non-invasive technique to measure brain activity and infer the cognitive state. With regard to audiovisual quality assessment, the study in [6] concluded that for longer sequences, low-quality conditions led to higher α and θ waves (result of EEG analysis), which respectively indicate decreased alertness and attention [1]. The α activity is also found to be significantly predictive of video quality [7]. In [8], a linear regression model based on α values and pupil dilation achieved a correlation value of 0.64 when predicting subjective QoE scores regarding video quality.

Other psychophysiological measures use electrocardiography (ECG), electrodermal activity (EDA) and eye measurement. In [9], ECG and EEG were found to be predictive of a 'Sensation of Reality' concerning perceived QoE for 2D and 3D multimedia stimuli. On the other hand, a direct relationship of EDA measurements with QoE could not be identified [6]. Eyes-related measurements may provide insight into cognitive activity relevant to QoE assessment that is not easily observable through other methods. For instance, eye movements provide valuable insight into overt visual attention whereas eye blink rate is related to visual fatigue and pupil dilation to cognitive load. 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 [10]. Furthermore, eye tracking data is often integrated into image and video quality metrics to further improve their quality prediction performance [11].

Differently from the aforementioned methods, QoE relationship with viewer's facial expressions has not been thoroughly investigated yet in literature, although it may have great potentials. Many studies can be found regarding facial expression recognition, (e.g., [12]), but these are just focused on the association of facial expressions extracted from face images to a specific emotion (e.g., happiness, sadness, anger). As a practical example, [13] presents a crowdsourcing web-based framework called Affectiva[2], which allows to collect and analyze facial expressions of video viewers to provide unobtrusive evaluation of facial responses to media content without relying on self-report ratings. However, the main objective is to determine the viewer's emotional engagement whereas the perceived quality is not considered. To the best of authors' knowledge, the studies in [2], [3] are the only two that propose to estimate the QoE on the basis of the viewer's facial expressions while watching on-screen video sequence. In [2], few details have been provided about the implementation of the emotion and quality prediction systems. Furthermore, the proposed model was trained using data obtained from only 3 subjects and validated by only 2 subjects. Also, the considered quality prediction is only aimed at identifying whether the video content is in line with viewer's content preferences. In [3], emotional factors are considered together with network QoS parameters to predict user's QoE. Different Machine Learning (ML) methods have been used to estimate the MOS in the basis of QoS and facial emotion parameters. The highest correlation (0.79) between subjective MOS and predicted MOS has been achieved with gradient based-boosting and Random Forest bagging based methods. However, the number of video (8) and testers (14) is limited for training and validation of the used ML proposed systems. Moreover, the proposed method estimates the MOS and not the subjective QoE perception of the single user.

With respect to these past works, we introduce the following novelties: i) we consider both facial expression and gaze direction to predict the QoE; ii) we have created a dataset made of a significant number of video sequences (400); iii) we have extensively experimented three different ML systems to analyze the potentialities of this psychophysiology-based approach for QoE estimation.

III. Overview of the proposed system

The objective of the proposed methodology is to investigate whether a correlation between the viewer's perceived QoE and the viewer's facial expression and gaze direction exists and to which extent this correlation may be helpful in the unobtrusive prediction of the viewer's QoE. Indeed, the past studies have demonstrated that human emotions may be derived from facial expressions and gaze direction may provide insights regarding viewer's visual attention. Based on this, we intend to go further by looking at the perceived quality.

Fig. 1 shows the flow chart for the proposed prediction system, which is composed of the following blocks:

1) *Data Acquisition*: a camera is used to detect and acquire the face of the person that is taking part to a video

[2] https://www.affectiva.com/experience-it/

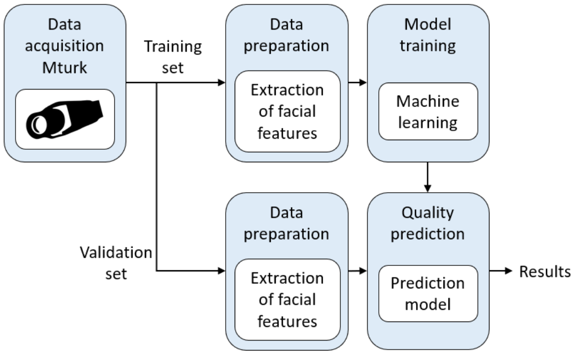

Fig. 1: Flow chart of the proposed system that shows the training and validation phases.

streaming session (i.e., the viewer). For our tests, the viewer's face is acquired by the PC's webcam of the participants of the crowdsourcing test implemented with the Amazon Mturk tool. At the end of each session the user is asked to rate the perceived video quality.

2) *Data Preparation*: facial features are detected and extracted from the face video to reduce the dimensionality of the information, identify specific facial expression information and remove any information which may identify the viewer (for privacy issues). Also, the viewer's gaze direction is analyzed to identify the visual attention and relevant features extracted and stored. Note that is performed for both the training and validation phases.
3) *Model training*: different machine learning algorithms are used for training the model on the basis of the data provided by the Data preparation block.
4) *Quality prediction*: a prediction model based on the trained machine learning classifier estimates the QoE perceived by the viewer on the basis of the analyzed facial expressions and gaze directions.

Observing the viewer's face expressions during the streaming session may indicate the degree of quality perceived as well as the presence of distortion events (e.g., low video quality, buffering). For example, if facial expressions reveal that the viewer is smiling, it may be supposed that the viewer is enjoying the video experience. Conversely, if facial expressions reveal a pout or annoyed expression, it is likely that the viewer is not happy about her experience, maybe because of insufficient video quality or frequent occurrence of annoying stalling events. Furthermore, the observation of the viewer's gaze direction may provide insights regarding the interest of the viewer to the watched video content.

Therefore, in this study we aim to implement a QoE prediction model able to predict the perceived viewer's QoE with regard to video streaming service just based on information extracted from the face of the viewer while watching the video and specifically, facial expressions and gaze direction. We want to highlight that privacy and security of viewers are safeguarded as the system only collects video features which do not contain any information related to the viewer identity.

IV. Dataset generation and model training

In this section we describe the procedure followed to create the dataset and to train the proposed system. The dataset is made available at the authors' lab portal.[3]

A. Crowdsourcing-based test

We conducted a crowdsourcing test to collect ground-truth quality perception values with reference to a video streaming service. We selected 5 different video contents from the LIVE Mobile Stall Video Database[4] [14], [15], from which we created 20 test video sequences. Specifically, we created 4 versions of the 5 original video contents by introducing different levels of initial delay and buffering events. The 5 original videos are: *Fantastic Finish Boys Basketball* (Basket), *Novak Djokovic vs Carlos Berlocq* (Novak), *Bike above the dust* (Bike), *Coldplay Paradise* (Coldplay) and *Football*.

The 4 versions of the test videos are as follows:

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

The test was completely developed using the crowdsourcing platform Amazon Mechanical Turk (MTurk). The whole test was implemented using the HTML5 markup language and the JavaScript language supported by the jQuery library. When a pretender participant selected the test from the MTurk platform, a first web page was shown describing the test and informing the pretender participant about the way the test should be conducted to receive the final reward, which was 1€. Privacy policies were also presented, highlighting that if the pretender participant agreed to participate to the test he/she automatically approved to be recorded with his/her webcam and allowed for the utilization of his/her video for research activities. If the pretender participant did not agree with these conditions, the test was interrupted. After the agreement of the participant with the test conditions, the web page containing the first video to watch appeared. When the participant was ready to watch the video, he/she started the video playing and his/her face was recorded during the watching. The video was shown in the center of the web page to facilitate the focus of the viewer on the video. When the video ended, it automatically disappeared and a banner appeared to notify the participant that the video of his/her face was correctly sent to the storage cloud space we used to store the recorded videos. Then, the participant was asked to rate the video quality

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

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

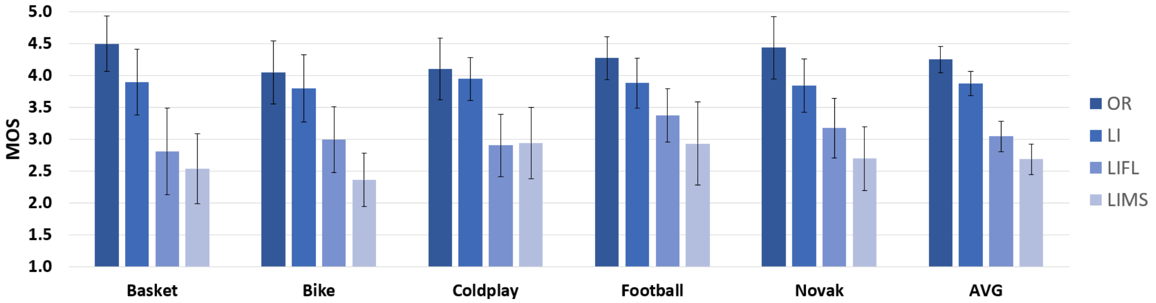

Fig. 2: MOS with 95% CI for each of the 20 videos and average MOS computed for each video version.

perceived choosing among five quality values as defined by ITU P.910 [16], i.e., 5 (Excellent), 4 (Good), 3 (Fair), 2 (Poor), 1 (Bad) [16]. The same procedure is repeated for all the 20 videos considered for the test. The time required to complete the test was approximately 25 minutes. Test participants were only required to have a HIT Approval Rate greater than 90%, no filtering was applied regarding age and/or country.

The test was completed by 20 participants and we collected a total of 400 quality values and relevant videos of the participant faces. In Fig. 2, we show the Mean Opinion Score (MOS) with 95% confidence interval (CI) computed for each of the 20 test videos and the average MOS computed for each video version (OR, LI, LIFL and LIMS). From these results it can be seen that for each different video content the original video with no impairments (OR) achieved the highest MOS value, followed respectively by the video with the long initial delay (LI), the video with long initial delay and few long buffering (LIFL) and finally the video with long initial delay and many short buffering (LIMS). Only for the Coldplay video content the MOS for the LIMS video is slightly higher than the MOS for the LIFL video.

B. *Features extraction*

The features we considered to train the prediction model are the facial expressions and gaze directions of the viewers while watching the video. By this, we aim to create a dataset which contains the sequence of the viewer's facial expressions during the video watching as well as his/her visual attention.

With regard to facial expression and recognition, common techniques used to extract relevant facial features from face images are Histogram of Gradient (HOG) and Local Binary Pattern (LBP) [17]. However, these approaches are based on the extraction and analysis of thousands of features for each single image (video frame), to which the proper emotion label is associated, which would introduce data size issues. Indeed, HOG and LBP are mostly used to compare different faces for people recognition purposes, as the data size of their features allow to precisely identify face's local patterns, such as the face's edges. Accordingly, we considered the Facial Action Coding System (FACS), which is the first codifier system for describing facial expressions by Action Units [18]. It uses static regular parameters to describe the contractions of specific facial muscles and the emotions related to these muscles. Each AU is labeled with a number and the description of muscle movement. By focusing on the muscle movements, we can limit the number of features while still keeping the meaningful information for our purposes.

The recognition and extraction of the AUs from the face image was done using two modules from the OpenFace project implemented by [19], [20]. Specifically, we considered 18 AUs to be extracted from each video frame, as these are the most relevant to describe human emotions [21]. For each video streaming session, we computed the average of each AU for all the video frames and the average movement of the muscle over its overall movement range. This allowed us to identify the sequence of movements of specific facial muscles of the viewer during the watching of the video as well as to reduce the features dataset to 36. Then, the first 18 features represent the average values of the 18 AUs, whereas the second sequence of 18 features represent the standard deviation of these AU.

With regard to the gaze direction, we extracted the position of the eyes' pupil using the same software used for extracting the AUs. We then considered the variance of the position of the eyes' pupils as features for our classifier over the whole sequence of frames of the viewer face video. From this, we were able to understand in which part of the video the visual attention of the viewer was focused or even to notice whether the viewer was not watching the video.

C. *QoE prediction model*

The objective of the QoE prediction model is to provide a prediction of the perceived viewer QoE on the basis of the viewer facial expressions and gaze directions during the watching of the video.

Specifically, we considered two prediction models:

- *AUtoQoE*: takes as input for each video the 40 features related to the viewer's facial expressions and as output the viewer's QoE;
- *AU&GAZEtoQoE*: takes as input for each video the 40 features related to the viewer's facial expressions and the position of the eyes' pupil related to the viewer's gaze direction, and as output the viewer's QoE.

Both the QoE prediction models were implemented with 3 different classifiers, namely Support Vector Machine (SVM) with quadratic kernel, RUSBoost trees, and bagged trees. The purpose was to find a pattern within the features dataset that could describe a correlation with the QoE final score provided by the viewers. The classifiers were implemented with the MATLAB software using the relevant machine learning libraries and the parallel computing toolbox supported by the CUDA drivers. The training/validation experiments followed a 5-fold cross-validation configuration to compute the final results in terms of accuracy, sensitivity and area under the curve (AUC), which are presented in Section V. With regard to QoE prediction, we considered three different quality scales:

- *5-level quality scale*: the QoE prediction model estimates the perceived QoE using the MOS scale from 1 to 5.
- *3-level quality scale*: the QoE prediction model estimates the perceived QoE using a 3-level quality scale where the first level includes the lowest 2 levels of the MOS scale (1-2), the second level is the third level of the MOS scale (3) and the third level includes the greatest 2 levels of the MOS scale (4-5). With this scale we aim to estimate whether the viewer is annoyed (1-2), sufficiently satisfied (3) or very satisfied (4-5).
- *2-level quality scale*: the QoE prediction model estimates the perceived QoE using a 2-level quality scale where the first level includes the lowest 2 levels of the MOS scale (1-2) whereas the second level includes the greatest 3 levels of the MOS scale (3-4-5). With this scale we aim to estimate whether the viewer is annoyed (1-2) or satisfied (3-4-5).

V. Results

To evaluate the performance of the proposed QoE prediction models, we considered the accuracy, the sensitivity, and the area under the curve (AUC) scores. From the results, which are shown in Table I, it can be seen that the classifiers perform better when trained with both the AU and gaze features than with only the AU features, in terms of all the considered performance metrics and for each of the 3 quality scales. This means that the information about the gaze direction of the viewer is directly connected to the perceived quality and allows to achieve a higher quality prediction accuracy. Furthermore, as was expected, the results achieved with the 2-level quality scale are better than those achieved with the 3-level quality scale, which in turn are better than those achieved with the 5-level quality scale. This means that it has not been found a strong correlation between the considered features and the 5-level MOS scores, meaning that a mapping between the emotions felt by the viewers regarding the video quality and the commonly used MOS scale is not the right choice in this case. This can be due to the fact that recognized facial expressions mainly identify positive or negative viewer's emotions 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 prediction as it may help to identify when the viewer is annoyed, which may be related to impairments affecting the video quality.

By comparing the performance of the classifiers, it results that the RUSBoost is the overall best classifier in terms of all the performance metrics. Indeed, although the RUSBoost does not achieve the greatest accuracy for the 2-level quality scale (but still comparable to the others), it achieves greater values of sensitivity, which are important as identify lower amount of false negatives. In particular, for the 2-level and 3-level quality scales, the RUSBoost achieves greater values of sensitivity with respect to SVM and Bagged trees, which are completely unreliable in the identification of the lowest quality values. The performance achieved by the classifiers are comparable for both the cases of training with AU features and AU&GAZE features. The reason why the RUSBoost achieved greater performance could be due to its capacity to learn better from imbalanced data. Indeed, using a combination of random under-sampling and boosting, RUSBoost is able to alleviate the class imbalance present among the collected MOS scores.

Finally, we compare the results of our best performing model, i.e., AU&GAZEtoQoE with RUSBoost classifier, with the models provided in [2] and [3] by considering the 5-level quality scale (i.e., the MOS scale). The models in [2] and [3] achieved a slightly greater Pearson Correlation Coefficient (PCC) than our model, respectively 0.82, 0.79 and 0.76. However, our model achieved higher accuracy (about 45%) than that achieved by the model in [2] (40%). In [3], the accuracy value was not provided. However, it must be highlighted that the model in [2] was trained and validated with data from 3 and 2 subjects, respectively, whereas 16 testers were involved for the model in [3] but it is not clear how many videos each tester had to watch. We trained the machine learning models with data from 400 videos therefore our results are more statistically relevant. Also, the model in [3] achieved the correlation of 0.79 training the machine learning model using both facial features and QoS features whereas the proposed model and the model in [2] are respectively based only on facial features and gaze directions, and facial features.

VI. Conclusion

We experimented three classifiers to predict the QoE perceived by an user of video streaming services analyzing her facial expression (through the action unit features) and gaze direction (through the position of the eye's pupil). We trained 3 different classifiers, namely SVM with quadratic kernel, RUSBoost trees and bagged trees with the considered features to predict the quality. We validated the QoE prediction models using three different quality scales: 5-level, 3-level and 2-level quality scales. The RUSBoost trees achieved the best results with the following accuracy: 44.7%, 59.4% and 75.3%using the 5-level, 3-level and 2-level quality scales. Furthermore, our model outperforms state-of-the-art models.

Whereas these results are not satisfactory yet, these represent a promising basis for further studies. Firstly, we intend to extend our experiment with other classifiers and to combine different classifiers together that may bring to better predictions. Secondly, we intend to further investigate on other features that may represent in a less accurate way the facial expressions but may be better in highlighting quality perception aspects. Thirdly, we intend to test unsupervised learning and compare the results with other psychophysiological methods.

TABLE I: Accuracy, sensitivity, and Area Under the Curve (AUC) score results obtained with the SVM, RUSBoost trees and Bagged trees classifiers for the AU&GAZEtoQoE and AUtoQoE prediction models.

Features	**Classifier**	**5-level quality scale**			**3-level quality scale**			**2-level quality scale**		
		AUC score	**Sensitivity**	**Acc.**	**AUC score**	**Sensitivity**	**Acc.**	**AUC score**	**Sensitivity**	**Acc.**
AU&GAZE	SVM	$1 = 0.73$ $2 = 0.66$ $3 = 0.61$ $4 = 0.62$ $5 = 0.80$	$1 = 0.33$ $2 = 0.23$ $3 = 0.40$ $4 = 0.36$ $5 = 0.72$	43.3%	$1/2 = 0.75$ $3 = 0.59$ $4/5 = 0.74$	$1/2 = 0.44$ $3 = 0.32$ $4/5 = 0.76$	56.2%	$1/2 = 0.75$ $3/4/5 = 0.75$	$1/2 = 0.37$ $3/4/5 = 0.89$	77.9%
	RUSBoost	$1 = 0.79$ $2 = 0.71$ $3 = 0.63$ $4 = 0.59$ $5 = 0.82$	$1 = 0.48$ $2 = 0.52$ $3 = 0.20$ $4 = 0.25$ $5 = 0.72$	44.7%	$1/2 = 0.75$ $3 = 0.63$ $4/5 = 0.75$	$1/2 = 0.72$ $3 = 0.32$ $4/5 = 0.63$	59.4%	$1/2 = 0.81$ $3/4/5 = 0.81$	$1/2 = 0.68$ $3/4/5 = 0.76$	75.3%
	Bagged	$1 = 0.79$ $2 = 0.71$ $3 = 0.64$ $4 = 0.62$ $5 = 0.81$	$1 = 0.33$ $2 = 0.31$ $3 = 0.44$ $4 = 0.30$ $5 = 0.70$	44.7%	$1/2 = 0.82$ $3 = 0.65$ $4/5 = 0.78$	$1/2 = 0.54$ $3 = 0.29$ $4/5 = 0.79$	59.0%	$1/2 = 0.82$ $3/4/5 = 0.82$	$1/2 = 0.35$ $3/4/5 = 0.93$	80.6%
AU	SVM	$1 = 0.76$ $2 = 0.65$ $3 = 0.60$ $4 = 0.61$ $5 = 0.78$	$1 = 0.30$ $2 = 0.26$ $3 = 0.41$ $4 = 0.31$ $5 = 0.68$	43.3%	$1/2 = 0.73$ $3 = 0.60$ $4/5 = 0.72$	$1/2 = 0.39$ $3 = 0.27$ $4/5 = 0.78$	55.8%	$1/2 = 0.69$ $3/4/5 = 0.69$	$1/2 = 0.35$ $3/4/5 = 0.90$	77.9%
	RUSBoost	$1 = 0.76$ $2 = 0.63$ $3 = 0.62$ $4 = 0.57$ $5 = 0.76$	$1 = 0.58$ $2 = 0.36$ $3 = 0.26$ $4 = 0.28$ $5 = 0.68$	41.9%	$1/2 = 0.73$ $3 = 0.61$ $4/5 = 0.72$	$1/2 = 0.60$ $3 = 0.38$ $4/5 = 0.57$	57.1%	$1/2 = 0.75$ $3/4/5 = 0.75$	$1/2 = 0.68$ $3/4/5 = 0.73$	71.9%
	Bagged	$1 = 0.82$ $2 = 0.71$ $3 = 0.57$ $4 = 0.50$ $5 = 0.79$	$1 = 0.30$ $2 = 0.28$ $3 = 0.34$ $4 = 0.21$ $5 = 0.65$	38.5%	$1/2 = 0.78$ $3 = 0.60$ $4/5 = 0.74$	$1/2 = 0.39$ $3 = 0.27$ $4/5 = 0.78$	55.5%	$1/2 = 0.78$ $3/4/5 = 0.60$	$1/2 = 0.28$ $3/4/5 = 0.93$	58.1%

References

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

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

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

[4] M. V. Peelen, A. P. Atkinson, and P. Vuilleumier, "Supramodal representations of perceived emotions in the human brain," *Journal of Neuroscience*, vol. 30, no. 30, pp. 10 127–10 134, 2010.

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

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

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

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

[9] E. Kroupi, P. Hanhart, J. Lee, M. Rerabek, and T. Ebrahimi, "Predicting subjective sensation of reality during multimedia consumption based on EEG and peripheral physiological signals," in *2014 IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2014, pp. 1–6.

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

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

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

[13] D. McDuff, R. el Kaliouby, and R. W. Picard, "Crowdsourcing facial responses to online videos: Extended abstract," in *Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 512–518.

[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)*, 2014, pp. 989–993.

[15] ——. (2016) LIVE Mobile Stall Video database. [Online]. Available: http://live.ece.utexas.edu/research/LIVEStallStudy/index.html

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

[17] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

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

[19] 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*, 2015, pp. 3756–3764.

[20] 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)*, vol. 06, 2015, pp. 1–6.

[21] S. Stöckli, M. Schulte-Mecklenbeck, S. Borer, and A. Samson, "Facial expression analysis with AFFDEX and FACET: A validation study," *Behavior Research Methods*, vol. 50, 2017.