Opinion Score Distribution Prediction via AI-Based Observers in Media Quality Assessment

Opinion Score Distribution Prediction via AI-Based Observers in Media Quality Assessment

Lohic Fotio Tiotsop, Antonio Servetti, Enrico Masala
Control and Computer Engineering Department
Politecnico di Torino
Turin, Italy
lohic.fotiotiotsop@polito.it, antonio.servetti@polito.it, enrico.masala@polito.it

***Abstract*—Training a Deep Neural Network (DNN) to predict an *individual*'s opinion score regarding the quality of multimedia content is a recent research direction. This type of DNN is called Artificial Intelligence-based Observer (AIO). By generating individual opinion scores, AIOs enable the prediction of the Opinion Score Distribution (OSD) for a given multimedia content. Multimedia image quality assessment literature lacks contributions that thoroughly assess the ability of AIOs to predict the OSD. In this paper a new set of AIOs is trained and shown to predict the OSD more accurately than state-of-the-art methods.**

***Index Terms*—Multimedia Quality Assessment, Opinion Score Distribution, Deep Neural Network, Users' Quality-of-Experience.**

I. Introduction

Multimedia quality assessment is a crucial task because it directly impacts users' quality-of-experience. High-quality multimedia content ensures user satisfaction and engagement, which are essential for the success of a multimedia platform. Moreover, an accurate quality assessment allows for the optimization of compression algorithms, bandwidth usage, and overall system performance, leading to a more efficient and effective delivery of multimedia content.

Typically, multimedia quality assessment is performed using objective and subjective methods. Objective methods involve computational algorithms that predict quality based on measurable system parameters (e.g., bit rate and frame rate) and content characteristics (e.g., assessing the presence of blurriness, blockiness, over/under exposure artifacts). These methods are fast but may not always correlate well with human perception. Subjective methods, on the other hand, involve human observers rating the quality of multimedia content through subjective experiments. These ratings are considered the gold standard, but subjective experiments are time-consuming and expensive to conduct.

Subjective experiments are typically conducted by asking a pool of subjects, usually more than 15, to rate the quality of multimedia content by choosing their opinion score on a scale ranging from 1 to 5, where 1 stands for bad quality and 5 corresponds to excellent quality. Due to varying expectations and experiences, subjects tend to provide different ratings, resulting in an Opinion Score Distribution (OSD). Figure 1 shows an example of OSD obtained from the ratings of 18 subjects.

While the multimedia quality assessment community has traditionally focused on predicting only the expected value of the OSD, known as the Mean Opinion Score (MOS), recent research has demonstrated fundamental advantages to go beyond the MOS [1]–[4]. It is now widely recognized that the entire OSD must be predicted to enable a more comprehensive assessment of multimedia quality [5]. This paper aims to advance the state-of-the-art on the OSD prediction.

Typical approaches to predicting the OSD make use of deep neural networks (DNNs). DNNs are trained to extract features from multimedia content and map them to the OSD, which serves as the ground truth in this context. This approach has been used for instance by the authors of [6]–[8]. Recently, other researchers have employed DNNs to train models capable of reproducing the multimedia quality assessments of an individual subject. These models are known as Artificial Intelligence-based Observers (AIOs) [9]–[12]. Many AIOs can be trained to mimic subjects with varying expectations, characteristics, and experiences. These AIOs can later be used to predict the quality of a multimedia content. Their combined predictions will generate different opinion scores, thereby providing an estimate of the OSD.

The ability of a pool of AIOs to predict the OSD compared to DNN-based models specifically tailored to OSD prediction has not been thoroughly studied within the multimedia quality assessment community. This paper addresses this issue. In particular, a new set of AIOs is trained. Unlike previous works focusing on AIOs, we consider a larger training set and employ a noise removal strategy that samples suitable subjects to model. The AIOs of the sampled subjects are trained using a convolutional DNN architecture strongly inspired by the well-known ResNet50 [13]. After the training process, the output of the softmax layer of the DNNs representing each AIO is used to derive a prediction of the OSD.

The results of computational experiments show that the set of trained AIOs can predict the OSD with a better accuracy than two state-of-the-art DNN-based models for OSD prediction as well as state-of-the-art AIOs. Furthermore, our results highlight the importance of opportunely sampling the subjects to be modeled, a step that has been disregarded in the previous literature on AIOs.

The rest of the paper is organized as follows. Section II briefly reviews the prior art, followed by Section III where

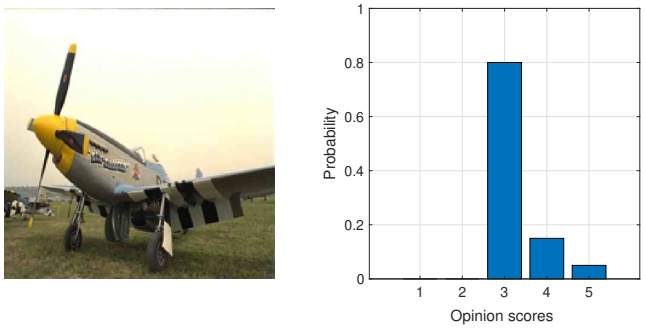


Fig. 1. The figure shows an example of OSD (right) for an image (left), i.e., the fraction of subjects that has selected each of the possible opinion scores when rating the quality of the image.

our approach to derive the OSD prediction from a set of newly trained AIOs is presented. Section IV discusses the results of computational experiments, while in Section V conclusions are drawn.

II. Literature Review

The research on multimedia quality assessment has long focused on designing algorithms to predict a single numerical value, i.e., the MOS [14]–[19]. These algorithms aim to predict the average quality as perceived by end-users. With the success of deep learning in various research fields, DNNs have also been naturally leveraged in the multimedia quality assessment community in an attempt to propose more accurate models for MOS prediction [20]–[23].

Despite the usefulness of algorithms for MOS prediction in several applications, the MOS alone is often insufficient for certain scenarios. A typical example is a video sequence containing artifacts in the background that are visible to some users but not to others. In this case, the MOS, as an average, is not a suitable indicator of overall quality. Such examples, along with many others highlighted in recent works, have motivated the necessity to go beyond the prediction of the MOS [1], [24], [25].

As a first step to go beyond the MOS, authors have suggested predicting, in addition to the MOS, the Standard Deviation of Opinion Score (SOS) [3], [4]. In this context, the SOS is seen as a measure of the disagreement among users' opinion scores. Other authors, however, have suggested that the subjectively perceived quality of multimedia content is a random variable; they recommended to predict the interval within which the quality lies with a specified probability [2].

In this dynamic shift towards predicting more comprehensive measures of quality than the MOS, the development of deep learning has also played a fundamental role. In fact, recently, DNN-based models to predict the entire OSD have been investigated. Clearly, the MOS, the SOS, as well as the interval in which the quality lies with a certain probability, can all be computed from the OSD. Thus, being able to predict the OSD not only allows for the preservation of previous metrics but also brings new possibilities. For instance, with the OSD, a content provider would know the percentage of subjects expected to be satisfied with the quality of any multimedia content coming from their production pipeline.

To predict the OSD, the authors of [6]–[8] prepared datasets in which each single image is associated with the related OSD obtained from a subjective experiment (see Figure 1). These subjectively annotated datasets were then used as a training set for DNN models aiming at OSD prediction. Thus, the OSD was considered as ground truth, and the DNNs were trained to extract suitable features from each image and map them to the OSD. Therefore this approach first aggregates the individual opinion scores collected from subjects with different characteristics to obtain the OSD. One shortcoming of this approach is that the trained DNN does not allow us to determine the characteristics of the subjects who are not satisfied with the quality of a given multimedia content. For this reason, researchers have recently suggested training a DNN, typically called an AIO, that can predict the opinion scores of a subject with well-defined characteristics [9]–[12]. In this way, the individual opinion scores predicted by many AIOs, modeling subjects with different characteristics, can later be aggregated as desired to obtain a prediction of the OSD or any other metric of interest.

The research on AIOs has so far mainly focused on two main questions: i) how to effectively train an AIO, since individual opinion scores are typically limited in size and also very noisy [10]; ii) what are the aspects of the subject scoring behavior that DNNs are able to mimic [9]. Therefore, the task of assessing to which extent AIOs are suitable for OSD prediction has not been thoroughly studied in the literature. This paper proposes an approach that integrates the strengths of a well-known model for noise removal in multimedia quality assessment with DNN-based modeling of individual opinion scores, i.e., AIOs, to achieve a more accurate prediction of the OSD.

III. OSD Prediction with AIOs

A. AIOs Training Process

Let us denote by $\mathcal{I}$ a set of subjects and by $\mathcal{J}$ a set of multimedia content. As mentioned in the introduction, during a subjective experiment, a subject $i \in \mathcal{I}$ is shown content $j \in \mathcal{J}$ and asked to select an opinion score from 1 to 5, representing their satisfaction with the quality of j. Generally, when the same content is shown to the same subject multiple times, the subject may not always select the same opinion score, introducing uncertainty in their choices. We will denote by p_{ijt} the probability that subject $i \in \mathcal{I}$ selects opinion score $t \in \{1, 2, 3, 4, 5\}$ when asked to rate content $j \in \mathcal{J}$.

The AIO of subject $i \in \mathcal{I}$, denoted by AIO_i is essentially a DNN trained to receive a multimedia content $j \in \mathcal{J}$ as input and predict as output the five probabilities p_{ijt}, where $t \in \{1, 2, 3, 4, 5\}$. To define the loss function and thus the optimization problem that guided the training process of our AIOs, let us denote by β_i the vector containing all the weights of AIO_i; by A_i the architecture of AIO_i and finally by $\hat{p}_{ijt}(\beta_i, A_i)$ the prediction of p_{ijt} made by AIO_i. The loss function used during the training process of AIO_i is the well

known multi-class cross entropy loss that is defined in our case as it follows:

$$L_i(\beta_i, A_i) = -\frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}}\sum_{t=1}^{5} p_{ijt} * \log(\hat{p}_{ijt}(\beta_i, A_i)); \quad (1)$$

where $|\mathcal{J}|$ denotes the cardinality of $\mathcal{J}$. The training of $AIO_i \ \ \forall i \in \mathcal{I}$ therefore consisted in finding the architecture A_i and the values of the related weights β_i so that $L_i(\beta_i, A_i)$ is minimized.

The architecture A_i for each $i \in \mathcal{I}$ was obtained by removing the fully connected and softmax layers from the original ResNet50 architecture and substituting them with a new fully connected layer with five neurons and a new softmax layer to predict a five-class probability distribution. Our choice for this architecture is motivated by previous research on AIOs, which has highlighted that suitable adjustments of the ResNet50 yield good baseline architectures for modeling individual subjects in terms of quality assessment [9], [10].

Given the architecture A_i, the weights β_i were trained by solving the following optimization problem.

$$\min_{\beta_i} L_i(\beta_i, A_i). \quad (2)$$

This problem was solved using the Stochastic Gradient Descent algorithm with Momentum (SGDM) [26]. In our implementation of SGDM, we initialized the weights β_i with those of the MDResNet50 model released by the authors of [9]. The learning rate was initially set to 10^{-2} and progressively decreased by dividing it by 10 every 5 epochs. The momentum parameter was set to 0.9 for the entire training process. The number of epochs varied from a minimum of 5 to a maximum of 50, depending on the subject.

We initialized our training process with the weights of the MDResNet50 rather than those of the ResNet50 because our work focuses on multimedia quality assessment, a task significantly different from image classification. In image classification, the model is designed to be robust against the presence of noise or compression artifacts in the input image, aiming to accurately identify objects despite such distortions. Conversely, in image quality assessment, a good model must be sensitive to these artifacts, as the task involves evaluating the quality of the images based on the presence and severity of such imperfections. Therefore, it is more appropriate to start with the weights of MDResNet50, a network trained specifically to characterize image artifacts, rather than ResNet50, which is optimized for image classification.

B. Training Set and Subject Selection

To train the AIOs, we used the SJTU IQSD dataset [27]. The SJTU IQSD dataset results from a subjective experiment involving 187 subjects. Each subject was shown a total of 808 images and asked to express their opinion on the quality. The 808 images used for the experiments were generated by adding different levels of blur, noise, JPEG, and JPEG2000 compression to 29 pristine quality images.

The SJTU IQSD dataset can therefore be used to train 187 AIOs, each corresponding to one subject. However, unlike previous works on AIOs that have modeled the entire set of subjects in the used dataset, in this work, we argue that selecting a suitable subset of subjects to model with a DNN is a fundamental step to enhance the accuracy of the final AIOs for the task they are trained for. Such a preliminary selection can also be seen as a noise removal step, since it is well known within the multimedia quality assessment community that individual opinion scores are significantly noisy. In fact, many approaches to recover the average subjectively perceived quality from raw opinion scores have been proposed [28]–[35]. Our strategy to identify a suitable subset of subjects to model relies on the subjective quality recovery model proposed in [30].

Denoting by r_{ij} the opinion score of subject i when asked to evaluate the quality of content j, the authors of [30] proposed the following model for the scoring behavior of a subject:

$$r_{ij} = q_j + b_i + N(0, \sigma_i), \quad (3)$$

where q_j represents the average subjectively percieved quality of content j, b_i and σ_i are the bias and inconsistency of subject i, respectively, and $N(0, \sigma_i)$ is a normally distributed noise term with mean equal to 0 and standard deviation equal to σ_i. Bias is intended here as a systematic tendency of a subject to provide lower (negative bias) or higher (positive bias) opinion scores with respect to the average perceived quality. Inconsistency, on the other hand, measures the subject's inability to consistently provide accurate opinion scores when evaluating media quality.

The model in Eq (3) is actually an ITU standardized tool in multimedia quality assessment to compute a noiseless estimate of the average subjectively perceived quality, i.e. q_j, from a given set of noisy raw opinion scores. Here, as we are interested in modelling individual ratings rather than the average perceived quality, we used this model for a different aim. In particular, using the set of ratings $\{r_{ij}\}$ collected during the SJTU ISQD experiment, we estimated, through maximum likelihood estimation, both the bias and the inconsistency of each of the 187 subjects. These bias and inconsistency values were used to determine a suitable set of subjects whose AIOs must be trained and used for the OSD prediction.

From the model in Eq (3), it can be seen that the higher the inconsistency of a subject, the less their opinion scores are determined by the actual quality of the images they are rating, since the noise term becomes particularly important. This lack of correlation means that it is impossible to train an accurate AIO to represent a highly inconsistent subject since there is no rational link between the image that the AIO would take as input and the rating it is expected to predict. In other words, an AIO trained on data from a highly inconsistent subject would not be able to learn a reliable mapping from image features to opinion scores, rendering it ineffective. Consequently, highly inconsistent subjects were excluded from the dataset to ensure the accuracy and reliability of the trained AIOs.

TABLE I
EARTH MOVER'S DISTANCE (EMD) BETWEEN THE GROUND TRUTH DISTRIBUTION OF OPINION SCORES AND THE PREDICTED ONE.

	LIVE R1S1	LIVER1S2	SJTU IQSD
NIMA [7]	0.289	0.239	0.233
***Gao et al* [6]**	0.206	0.213	0.109
AIOs from [10]	0.369	0.372	0.300
AIOs from [9]	0.285	0.312	0.254
Proposed AIOs	**0.159**	**0.161**	**0.100**

TABLE II
ROOT MEAN SQUARE ERROR (RMSE) BETWEEN THE GROUND TRUTH DISTRIBUTION OF OPINION SCORES AND THE PREDICTED ONE.

	LIVE R1S1	LIVER1S2	SJTU IQSD
NIMA [7]	0.285	0.219	0.182
***Gao et al* [6]**	0.234	0.242	**0.089**
AIOs from [10]	0.355	0.297	0.269
AIOs from [9]	0.308	0.262	0.219
Proposed AIOs	**0.197**	**0.146**	0.092

On the other hand, we could not only keep the subject with the lowest inconsistency and exclude all the other subjects as we needed to retain enough subjects to represent different expectations and characteristics in order to accurately predict the OSD. To achieve this, we needed to establish a threshold for inconsistency, above which subjects are excluded from the dataset. Simultaneously, we had to ensure that the selected subjects have varying levels of expectations, as represented by their bias. This approach ensures that the dataset remains diverse enough to capture different expectations and preferences, while also maintaining the rationality and consistency necessary for training effective AIOs.

We empirically tested different values as the threshold for inconsistency. For each value of the threshold, we identified all the subjects with inconsistency less than or equal to that threshold. We then selected a subset of these subjects such that their bias values were almost uniformly distributed between the maximum and minimum bias values observed in the dataset in order to represent different levels of expectations. We trained the AIOs for the selected subjects, used them to predict the OSD, and recorded their prediction accuracy. We observed that the best prediction accuracy was achieved by setting the threshold to 0.5. With this threshold, a set $\mathcal{I}_{selected}$ containing a total of 45 subjects was identified. The final OSD prediction was then derived from the 45 AIOs of the subjects in $\mathcal{I}_{selected}$ as explained in the next section.

C. From AIOs to the OSD prediction

We remind that given a multimedia content j, we are interested in predicting the fraction f_{jt} of end users that would score the quality of j by selecting t as their opinion score, with $t \in \{1, 2, 3, 4, 5\}$. As already mentioned, once the AIO of a subject i, i.e., AIO_i, is trained, it receives the content j as input and outputs the probabilities $p_{ijt}(\beta_i, A_i)$, from which we derive a prediction of the desired fraction f_{jt} as follows:

$$\hat{f}_{jt} = \frac{1}{|\mathcal{I}_{selected}|} \sum_{i \in \mathcal{I}_{selected}} p_{ijt}(\beta_i, A_i). \quad (4)$$

Please note that at the inference time, both the architecture A and the weights β are known for all the AIOs, and thus $\hat{f}_{jt}$ is simply a numerical value. By putting together the five values $\{\hat{f}_{jt} \quad t = 1, 2, \ldots, 5\}$, one obtains the desired prediction of the OSD.

IV. RESULTS

In this section, the results of computational experiments conducted to assess the effectiveness of our proposal are presented and discussed. The performance of the trained AIOs in terms of accuracy in the OSD prediction were tested on two datasets never used during the training phase, i.e. the first and the second sessions of the first release of the LIVE image quality assessment dataset [36] (LIVE R1S1 and LIVE R1S2). For completeness' sake, the AIOs performance on the SJTU IQSD dataset, used as training set, was also reported.

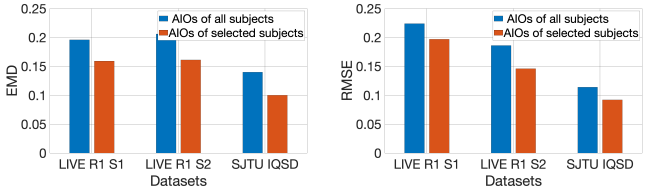
A. Quantitative analysis

In this section we compare the performance of the 45 trained AIOs of the selected sample of subjects to that of two DNN-based approaches for OSD prediction, i.e., NIMA [7] and the model proposed by *Gao et al.* [6], as well as that of the AIOs published in [10] and in [9]. The comparison was made in terms of Earth Mover's Distance (EMD) [6] and Root Mean Square Error (RMSE) between the predicted OSD and the ground truth.

All the different models were used to predict the OSD on the LIVE R1S1, LIVE R2S2 and SJTU IQSD datasets. The average RMSE and EMD were then calculated for each model on each dataset. The results for the EMD and RMSE are reported in Table I and Table II, respectively. The lower the EMD and the RMSE, the better the model.

It can be seen from Table I and Table II, the trained AIOs generally predicted the OSD with greater accuracy. In fact, in five out of six testing conditions, the proposed set of AIOs provided an OSD prediction with a lower average EMD and RMSE compared to the ground truth. Only in the case of the SJTU IQSD dataset, the model from *Gao et al.* slightly outperformed the proposed AIOs in terms of RMSE. It is worth mentioning that the model from *Gao et al.* used here was trained on the SJTU IQSD dataset. It is therefore natural that this model performs quite well on that dataset.

When comparing the performance of the proposed AIOs to that of state-of-the-art AIOs in terms of OSD prediction in Table I and Table II, it can be noticed that the proposed AIOs



(a) EMD (b) RMSE

Fig. 2. Comparison of the OSD prediction accuracy of the AIOs of the selected subjects to that of all subjects.

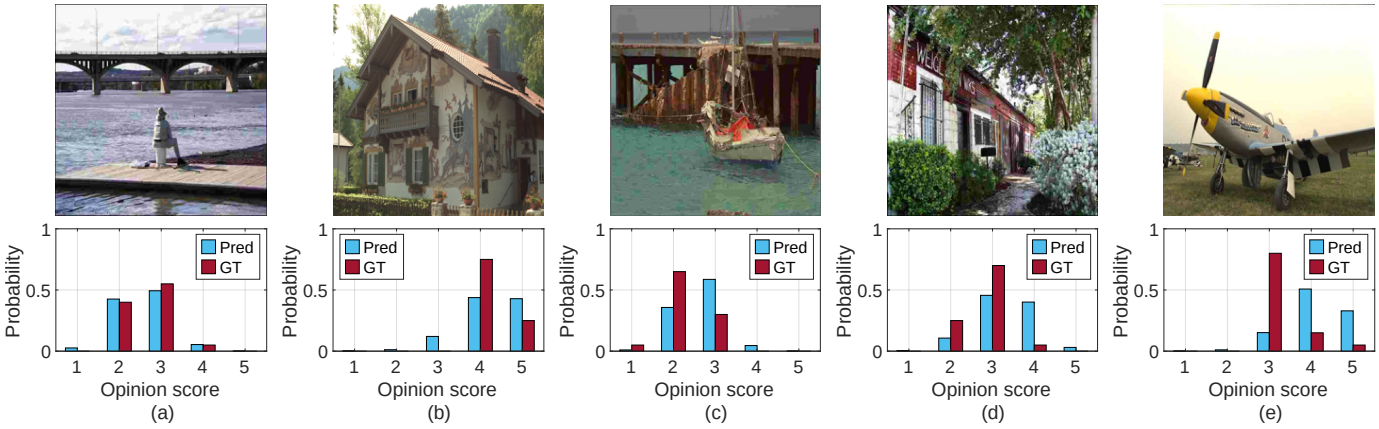


Fig. 3. Visual comparison between the OSD predicted by the proposed AIOs (Pred) and the ground truth (GT) for five images from the LIVE R1S1 dataset. The five images are presented in decreasing order of OSD prediction accuracy, measured in terms of EMD between the predicted OSD and the GT.

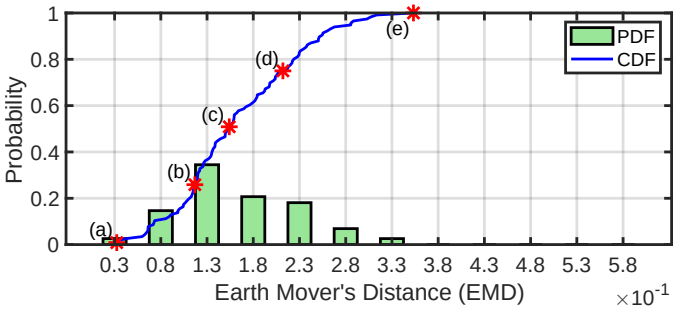


Fig. 4. Distribution (PDF) and cumulative distribution (CDF) of Earth Mover's distance (EMD) values for all the images in the LIVE R1S1 dataset.

performed significantly better. This can be explained by two facts: i) we are modeling a larger set of subjects (45 compared to the 19 considered in [10] and [9]), thus considering a wider range of subject expectations and preferences; ii) we are implementing a noise removal strategy by excluding highly inconsistent subjects, a step overlooked in [10] and [9].

B. Assessing the Importance of Subject Selection

In order to illustrate the relevance of the subject selection carried out in Section III-B, we compared the performance of the AIOs of all 187 subjects to that of the 45 selected subjects. The results are summarized in Figure 2a and Figure 2b, for the EMD and the RMSE, respectively.

As it can be seen from Figure 2a and Figure 2b, using the AIOs of all subjects to predict the OSD resulted in lower accuracy compared to using only the AIOs of the selected subjects. In fact, on all three datasets, the predictions from the selected set of subjects yielded a lower average EMD and RMSE. This highlights the importance of our subject selection approach prior to the training of the AIOs.

It is particularly interesting to note that even on the training set, the selected subset of subjects performed better than the set of all subjects in predicting the distribution of all opinion scores (including those of non-selected subjects). This suggests, as mentioned earlier, that a DNN can hardly learn any relevant information from the opinion scores of highly inconsistent subjects.

C. Qualitative analysis

Figure 3 shows a visual comparison between the predicted OSD by the proposed AIOs and the ground truth for five images from the LIVE R1S1 dataset. The five images are presented in decreasing order of OSD prediction accuracy, measured by the EMD. In particular, the image labeled as (a) has the highest prediction accuracy (minimum EMD), while the images labeled as (b), (c), and (d) correspond to the 25th, 50th, and 75th percentiles of the EMD values distribution, respectively. Finally, the image labeled as (e) corresponds to the worst prediction, i.e., the one with the maximum EMD. Figure 4 shows the cumulative distribution of the EMD values for all the images in the LIVE R1S1 dataset, which can be used to identify the EMD values corresponding to the aforementioned percentiles. For instance, the 25th percentile is around $1.11 * 10^{-1} = 0.11$.

From Figure 3, it can be seen that even in the worst case (labeled as (e)), the support of the predicted distribution (Pred) significantly intersects that of the ground truth distribution (GT). Additionally, in all cases, the mode of the predicted distribution differs from that of the ground truth by at most one quality level. This suggests that there are no cases where there is a significant disagreement between the subjects' opinion scores and those of the proposed AIOs.

It is also worth noting that in all cases, the predicted distribution displayed a shape typically expected from subjects in a quality-of-experience experiment. Empirically, one expects a unimodal distribution with all selected opinion scores concentrated around the mode, representing the opinion of the majority within the population. The distribution predicted by the proposed AIOs satisfied these empirical properties in all cases from the best to the worst prediction.

V. Conclusion

This paper focused on the task of predicting the OSD in multimedia quality assessment. Using a subjectively annotated

dataset containing the ratings of 187 subjects, we presented an approach to select 45 of these subjects whose ratings can be suitably modeled with a DNN. We then trained a DNN for each of the selected subjects, yielding 45 AIOs representing subjects with different levels of expectations. The trained AIOs were then used to predict the OSD on several datasets. The results highlighted that the proposed AIOs predict the OSD with greater accuracy than several state-of-the-art approaches.

REFERENCES

[1] T. Hoßfeld, P. E. Heegaard, M. Varela, and S. Möller, "QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to mos," *Quality and User Experience*, vol. 1, no. 1, Sep 2016.

[2] L. Fotio Tiotsop, E. Masala, A. Aldahdooh, G. V. Wallendael, and M. Barkowsky, "Computing quality-of-experience ranges for video quality estimation," in *11th Intl. Conference on Quality of Multimedia Experience (QoMEX)*. Berlin, Germany: IEEE, Jun 2019, pp. 1–3.

[3] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!" in *Third International Workshop on Quality of Multimedia Experience (QoMEX)*. Mechelen, Belgium: IEEE, Sep 2011, pp. 131–136.

[4] L. F. Tiotsop, T. Mizdos, M. Uhrina, M. Barkowsky, P. Pocta, and E. Masala, "Modeling and estimating the subjects' diversity of opinions in video quality assessment: a neural network based approach," *Multimedia Tools and Applications*, vol. 80, pp. 1–19, 2020.

[5] M. Seufert, "Fundamental advantages of considering quality of experience distributions over mean opinion scores," in *2019 11th International Conference on Quality of Multimedia Experience (QoMEX)*. Berlin, Germany: IEEE, June 2019, pp. 1–6.

[6] Y. Gao, X. Min, Y. Zhu, J. Li, X.-P. Zhang, and G. Zhai, "Image quality assessment: From mean opinion score to opinion score distribution," in *Proceedings of the 30th ACM International Conference on Multimedia*. Lisbon, Portugal: ACM, 2022, pp. 997–1005.

[7] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[8] D. Varga, D. Saupe, and T. Szirányi, "Deeprn: A content preserving deep architecture for blind image quality assessment," in *2018 IEEE Intl. Conference on Multimedia and Expo (ICME)*, 2018, pp. 1–6.

[9] L. Fotio Tiotsop, A. Servetti, P. Pocta, G. Van Wallendael, M. Barkowsky, and E. Masala, "Multiple image distortion DNN modeling individual subject quality assessment," *ACM Transactions on Multimedia Computing, Communications and Applications*, May 2024.

[10] L. F. Tiotsop, A. Servetti, M. Barkowsky, P. Pocta, T. Mizdos, G. Van Wallendael, and E. Masala, "Predicting individual quality ratings of compressed images through deep CNNs-based artificial observers," *Signal Processing: Image Communication*, vol. 112, p. 116917, Mar. 2023.

[11] L. F. Tiotsop, T. Mizdos, M. Barkowsky, P. Pocta, A. Servetti, and E. Masala, "Mimicking individual media quality perception with neural network based artificial observers," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 18, no. 1, p. 1–25, 2022.

[12] L. F. Tiotsop, T. Mizdos, M. Uhrina, P. Pocta, M. Barkowsky, and E. Masala, "Predicting single observer's votes from objective measures using neural networks," in *Human Vision and Electronic Imaging conference (HVEI)*. San Francisco, USA: Society for Imaging Science and Technology (IS&T), Jan 2020, pp. 130–1 – 130–8,.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE conference on computer vision and pattern recognition*. Las Vegas, NV, USA: IEEE, 2016, pp. 770–778.

[14] A. K. Vishwakarma and K. M. Bhurchandi, "No-reference video quality assessment using novel hybrid features and two-stage hybrid regression for score level fusion," *Journal of Visual Communication and Image Representation*, vol. 89, p. 103676, 2022.

[15] D. Pan, X. Wang, P. Shi, and S. Yu, "No-reference video quality assessment based on modeling temporal-memory effects," *Displays*, vol. 70, p. 102075, 2021.

[16] L. F. Tiotsop, A. Servetti, and E. Masala, "Full reference video quality measures improvement using neural networks," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, 2020, pp. 2737–2741.

[17] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2256–2270, 2018.

[18] L. F. Tiotsop, T. Mizdos, E. Masala, M. Barkowsky, and P. Pocta, "How to train no reference video quality measures for new coding standards using existing annotated datasets?" in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*. Tampere, Finland: IEEE, 2021, pp. 1–6.

[19] L. F. Tiotsop, F. Agboma, G. Van Wallendael, A. Aldahdooh, S. Bosse, L. Janowski, M. Barkowsky, and E. Masala, "On the link between subjective score prediction and disagreement of video quality metrics," *IEEE Access*, vol. 9, pp. 152 923–152 937, 2021.

[20] X. Liu, J. van de Weijer, and A. D. Bagdanov, "Rankiqa: Learning from rankings for no-reference image quality assessment," in *The IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy: IEEE, Oct 2017, pp. 1040–1049.

[21] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, Washington: IEEE, 2020, pp. 3572–3582.

[22] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2018.

[23] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, USA: IEEE, 2020, pp. 3664–3673.

[24] K. Mitra, A. Zaslavsky, and C. Ahlund, "Context-aware QoE modelling, measurement and prediction in mobile computing systems," *IEEE Transactions on Mobile Computing*, vol. 14, pp. 920–936, May 2015.

[25] S. Pezzulli, M. G. Martini, and N. Barman, "Estimation of quality scores from subjective tests: beyond subjects' MOS," *IEEE Transactions on Multimedia*, vol. 23, pp. 2505–2519, 2020.

[26] Y. Liu, Y. Gao, and W. Yin, "An improved analysis of stochastic gradient descent with momentum," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 261–18 271, 2020.

[27] Y. Gao, X. Min, W. Zhu, X.-P. Zhang, and G. Zhai, "Image quality score distribution prediction via alpha stable model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 6, pp. 2656–2671, 2023.

[28] L. Janowski and M. Pinson, "The Accuracy of Subjects in a Quality Experiment: A Theoretical Subject Model," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2210–2224, Dec. 2015.

[29] Z. Li and C. G. Bampis, "Recover Subjective Quality Scores from Noisy Measurements," in *Proc. Data Compression Conference (DCC)*. Snowbird, UT, USA: IEEE, Apr. 2017, pp. 52–61.

[30] Z. Li, C. G. Bampis, L. Janowski, and I. Katsavounidis, "A Simple Model for Subject Behavior in Subjective Experiments," *Proc. Intl. Symposium on Electronic Imaging*, vol. 32, no. 11, pp. 131–1–131–14, Jan. 2020.

[31] J. Li, S. Ling, J. Wang, and P. Le Callet, "A Probabilistic Graphical Model for Analyzing the Subjective Visual Quality Assessment Data from Crowdsourcing," in *Proc. 28th Intl. Conf. on Multimedia*. Seattle, WA, USA: ACM, Oct. 2020, pp. 3339–3347.

[32] L. Fotio Tiotsop, A. Servetti, M. Barkowsky, and E. Masala, "Modeling subject scoring behaviors in subjective experiments based on a discrete quality scale," *IEEE Transactions on Multimedia*, pp. 1–16, Mar. 2024.

[33] L. F. Tiotsop, A. Servetti, M. Barkowsky, and E. Masala, "Regularized maximum likelihood estimation of the subjective quality from noisy individual ratings," in *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2022, pp. 1–4.

[34] L. Fotio Tiotsop, A. Servetti, and E. Masala, "A scoring model considering the variability of subjects' characteristics in subjective experiments," in *Proc. 15th Intl. Conf. on Quality of Multimedia Experience (QoMEX)*. Ghent, Belgium: IEEE, Jun. 2023, pp. 1–6.

[35] A. Altieri, L. F. Tiotsop, and G. Valenzise, "Subjective media quality recovery from noisy raw opinion scores: A non-parametric perspective," *IEEE Transactions on Multimedia*, pp. 1–16, 2024.

[36] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "Live image quality assessment database," *http://live.ece.utexas.edu/research/quality*, vol. 1, p. 1, 2005.