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

Lohic Fotio Tiotsop

VQEG JEG-Hybrid

Nov 21, 2024

Outline

1. From Mean Opinion Score (MOS) to Opinion Scores Distribution (OSD)
2. OSD prediction with Artificial Intelligence-based Observers
3. Results and Discussion
4. Conclusions

From MOS to Opinion Score Distribution (OSD)

- **Opinion Score Distribution (OSD):** represents the full distribution of individual ratings for an image/video, showing how many evaluators chose each score (e.g., 1 to 5)
- **Why OSD over MOS?** recent research has highlighted the advantages of OSD:
  - Captures the variability and disagreements
  - Offers a deeper understanding of users' satisfaction

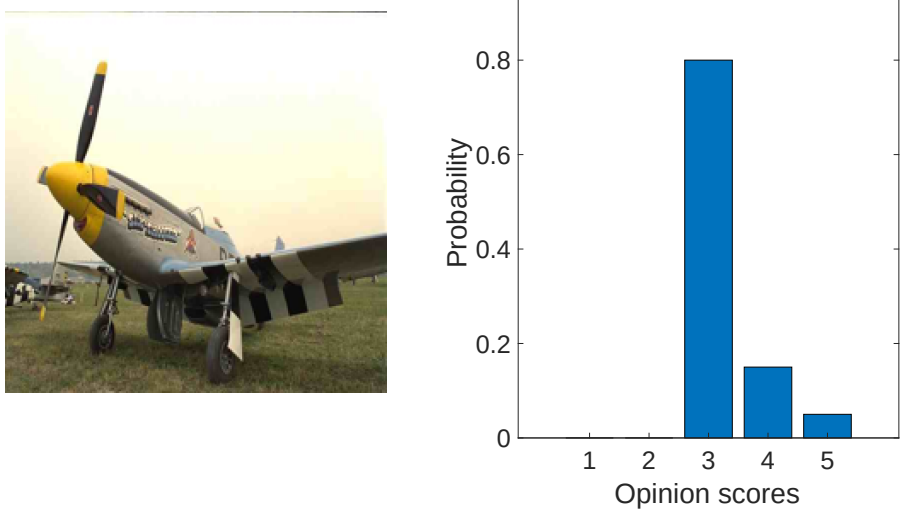


Figure: Example of OSD (right) for an image (left)

Opinion Score Distribution (OSD) Prediction Task

- **Modern Approach**
  - Individual opinion scores in the training set are first aggregated to obtain the OSD
  - A single Deep Neural network (DNN) is then trained to predict the obtained OSD
- **Our Approach**
  1. Select, from the training set, a set of accurate subjects with diverse levels of expectation in order to best represent the OSD
  2. Train as many DNNs as the number of selected subjects, each mimicking one subject
  3. Use the trained DNNs to predict several individual opinion scores and thus the OSD
- **Note**: we refer to a DNN trained to predict individual opinion scores as an "Artificial Intelligence-based Observer" (AIOs)

Training Set and Subject Selection

- **Training set:** SJTU IQSD dataset, 187 subjects, each rating 808 images with different distortions.
- **Subject Selection:** A subset of subjects was chosen to enhance AIO accuracy by reducing noise.
- We used the **SUREAL scoring model** for subject selection:

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

where r_{ij} is the opinion score of subject i for image j, q_j is the perceived quality of image j, b_i and σ_i are the bias and the inconsistency of subject i.

- **The inconsistency represents the subject's level of accuracy, while the bias captures the subject's level of expectation**

Subject Selection Strategy

- **Inconsistency Threshold:** Subjects with $\sigma_i > 0.5$ were excluded to ensure models' reliability.
  - It was experimentally found that there is no advantages in further decreasing this threshold.
- **Diversity:** we selected the subjects to represent varying biases to capture different expectations.
  - Subjects selected so that their bias values were uniformly distributed between the min and max estimated bias values
- **Result:** 45 subjects, out of the 187 available ones, were selected and we trained their corresponding AIOs

AIOs Training Process

- **Objective:** to train the AIO of subject i (AIO_i)
- **Input/Output:** AIO_i receives image $j \in \mathcal{J}$ and output probabilities $\hat{p}_{ijt}$ for scores $t \in \{1, 2, 3, 4, 5\}$.
- $\hat{p}_{ijt}$ is the predicted probability that subject i scores the quality of image j as t.
- **Loss Function:** multi-class cross-entropy loss for the training of AIO_i:

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

where $\hat{p}_{ijt}$ is the true probability obtained from a one hot encoding of the opinion scores of subject i

AIOs Architecture and Training

- **Architecture:** ResNet50 with modified fully connected and softmax layers for five-class probability distribution.
- **Weights initialisation:** the training process was initialized with the weights of the MDResNet50.
- **The MDResNet50:** A pre-trained deep CNN, to classify images based on distortion type and level, making it a better starting point than ResNet50 for IQA tasks.
- **Training Algorithm and Parameters:**. stochastic gradient descent algorithm with momentum; initial learning rate set to 10^{-2}, multiplied by 0.1 every 5 epochs; momentum set to 0.9.

From AIOs to OSD Prediction

- **Goal:** Predict the fraction f_{jt} of users assigning score t to content j.
- **Inference:** Calculate $\hat{f}_{jt}$ from AIO predictions as:

$$\hat{f}_{jt} = \frac{1}{|\mathcal{I}_{selected}|} \sum_{i \in \mathcal{I}_{selected}} \hat{p}_{ijt}$$

  where $\mathcal{I}_{selected}$ represents the set of selected and modeled subjects
- **Result:** Aggregate $\{\hat{f}_{jt}\}$ for $t = 1, 2, ..., 5$ to obtain the predicted OSD for image j.

Results: EMD Comparison

- **Objective:** to compare Earth Mover's Distance (EMD) between predicted and ground truth OSD.
- **Datasets:** LIVE R1S1, LIVE R1S2, SJTU IQSD.
- **Results:**
  - **The proposed AIOs** provided the lowest EMD on all datasets (0.159, 0.161, 0.100).
  - The proposed AIOs outperformed existing ones that ignored the subject selection step for noise removal

Table: EMD values from different approaches

	LIVE R1S1	LIVER1S2	SJTU IQSD
NIMA [1]	0.289	0.239	0.233
Gao et al [2]	0.206	0.213	0.109
AIOs from [3]	0.369	0.372	0.300
AIOs from [4]	0.285	0.312	0.254
Proposed AIOs	**0.159**	**0.161**	**0.100**

Results: RMSE Comparison

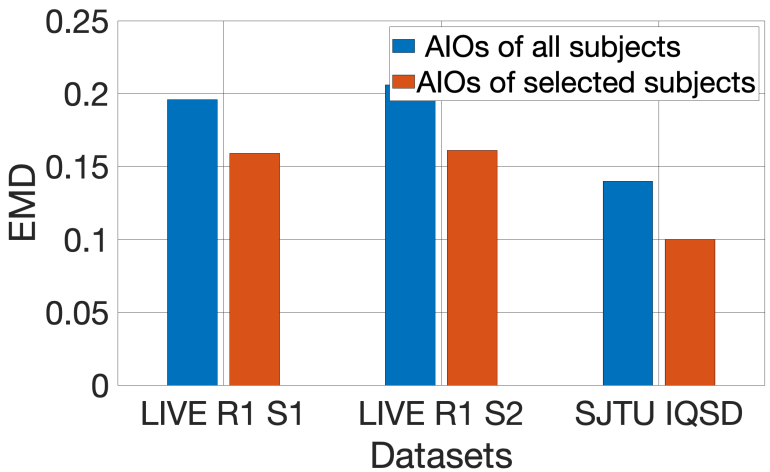
- **Objective:** to compare Root Mean Square Error (RMSE) between predicted and ground truth OSD.
- **Datasets:** LIVE R1S1, LIVE R1S2, SJTU IQSD.
- **Results:**
  - **Proposed AIOs:** Lowest RMSE on LIVE R1S1 (0.197) and LIVE R1S2 (0.146).
  - **Other Models:** Gao et al. slightly better on SJTU IQSD (0.089 vs. 0.092).

Table: RMSE values from different approaches

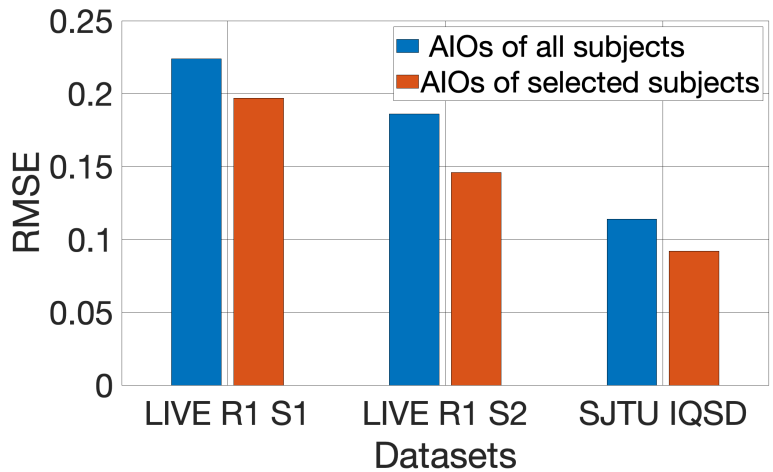
	LIVE R1S1	**LIVER1S2**	**SJTU IQSD**
NIMA [1]	0.285	0.219	0.182
***Gao et al* [2]**	0.234	0.242	**0.089**
AIOs from [3]	0.355	0.297	0.269
AIOs from [4]	0.308	0.262	0.219
Proposed AIOs	**0.197**	**0.146**	0.092

Results: Importance of Subject Selection

- **Comparison:** All subjects vs. selected 45 subjects.
- **Results:** Selected subjects provided better predictions in terms of EMD and RMSE.



(a) EMD



(b) RMSE

Figure: Comparison of the OSD prediction accuracy of the AIOs of the selected subjects to that of all subjects.

Results: Qualitative Analysis

- **Visual Comparison:** OSD predictions vs. ground truth for five images.
- **Observation:**
  - The predicted distributions generally align with ground truth.
  - Even in the worst case (e), the predicted distributions shares the same support with the ground-truth.

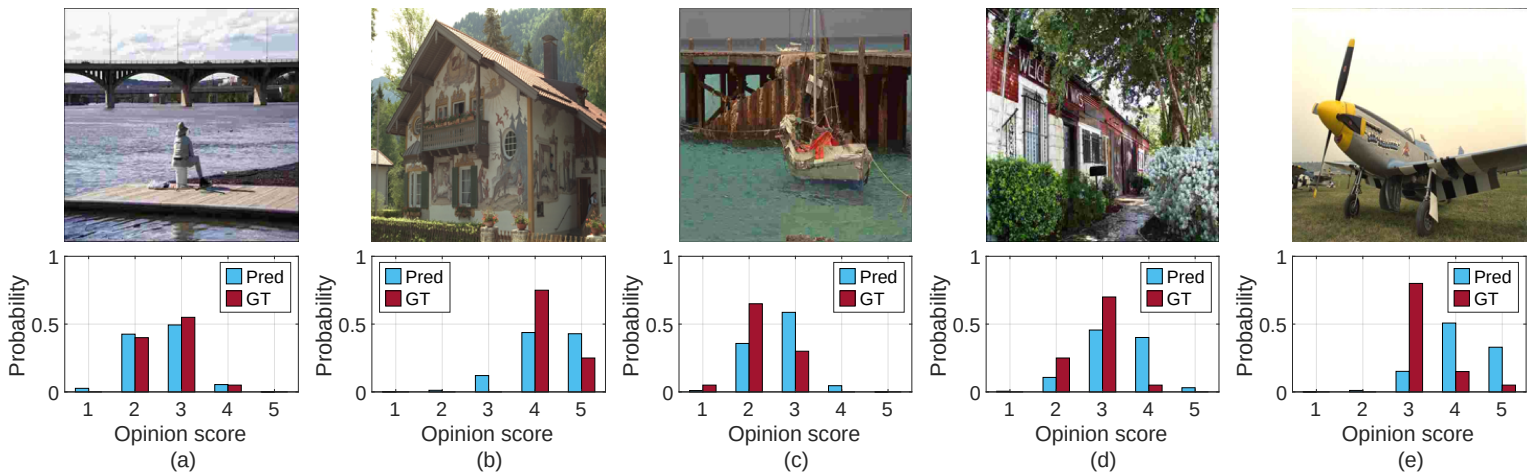


Figure: Visual comparison of OSD predictions (Pred) and ground truth (GT) for five images from the LIVE R1S1 dataset, ranked by decreasing prediction accuracy based on EMD

Conclusion

- **Superior Accuracy:** Proposed AIOs outperformed state-of-the-art models with lower EMD and RMSE.
- **Subject Selection:** Improved accuracy by removing inconsistent subjects.
- **Robust Performance:** Consistent alignment with ground truth in all cases.
- **Future Work:** Enhance AIOs performance by exploring more task-inspired loss functions and model architectures

References

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

[2] 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.

[3] 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.

[4] 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.

Thanks for your attention