How to Evaluate a Biometric Facial Image

Diana Gonçalves D'Amil

2022231368

Undergraduate Project in Biomedical Engineering

Advisor:

Professor Nuno Gonçalves

FCTUC, Physics Department

Institute of Systems and Robotics, Coimbra, Portugal

Coimbra, 2025

Table of Contents

List of Figures

Introduction

FIQA or Face Image Quality Assessment is the process of evaluating the quality of a facial image by considering factors such as sharpness, lighting, pose, and facial expressions to ensure efficient performance in facial recognition systems. (Athar et al.) This evaluation is important for improving the performance of systems such as facial recognition, biometric authentication and medical imaging. The consequences of poor-quality images extend beyond performance drops. For example, in medical imaging, inaccurate interpretation can impact diagnostic decisions. Ensuring that images meet a perceptual quality threshold is therefore essential for accuracy and security.

While automated systems increasingly rely on facial imagery, there remains a gap between objective quality assessment and human perception. This gap can lead to system failures, highlighting the urgent need for more human-aligned quality metrics.

This project investigates whether fusion-based objective metrics can better approximate subjective facial image quality compared to traditional standalone measures. For that, we use a dataset of facial images with controlled distortions and collect subjective quality scores from human evaluation. We then apply various objective image quality assessment metrics and test them, and the fusion of them through a set of approaches, to see which of them better approximates to subjective perception. This has the goal to bridge the gap between objective evaluation and human perception and judgement.

Most of the facial images we have access to may have reduced quality, filters, or distortions that make facial recognition difficult. Image quality assessment allows us to improve facial recognition by eliminating images that are considered poor in quality based on relative human, while also helping to improve facial recognition systems by enabling them to better handle low-quality inputs. A low-quality image, or low FIQ, is considered such due to characteristics like poor lighting, pose variation, misalignment, blur, among others.

There are two main ways to classify a facial image: objective FIQ (OFIQ) and relative FIQ (RFIQ). OFIQ considers the previously mentioned characteristics, while

RFIQ assesses the difference between the presented image and the original and high-quality one, in other words, the evaluation is based on how much the given image differs in quality from the original. (Kim et al.). Importantly, RFIQ is inherently subjective, as it depends on each person's perception of what constitutes a degradation in quality.

The OFIQ can be classified as Full-Reference (FR), Reduced-Reference (RR) and No-Reference (NR). In the first the original image is used as a reference, but the assessment relies on computational metrics to quantify the similarity or difference, rather than subjective perception. This makes FR a more objective evaluation method compared to RFIQ. This is only appropriate when the reference image is of very good quality. RR methods have access only to certain features of the original image. The NR IQA consists of an evaluation where it has no access to the reference image and only estimates the image quality. The latter is more used despite being limited because it relies on assumptions about typical distortion patterns and statistical models of natural images. (Athar et al.)

The more used objective metrics are the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) along with its many derivatives. They provide automated evaluations of image quality, yet there's an inconsistency between this assessment and the subjective one, which can be problematic in facial images, mainly because the perception is an important part of the evaluation (Wang et al.).

In subjective quality assessment, humans evaluate the visual quality of the image, assigning them quality scores (QS) (T. Schlett et al.) and the average of this opinions is referred to as MOS or Mean Opinion Score. Since subjective assessment is the closest representation of general human opinion, the goal of objective assessment is to approximate it, that is, to enable algorithms to predict image quality similarly to how humans would, because the subjective assessment is time-consuming and high-priced.

However, human opinion is quite difficult to predict, since the evaluation of faces, unlike of general objects, by the human brain is processed in a specific region called the fusiform face area, which is highly sensitive to subtle visual cues. Perception of faces quality is influenced by various demographic and non-demographic factors. This explains how perception is receptive to stimulus such as age, gender, attractiveness and ethnicity

(Tsao et al. and Neto et al.). This complexity shows the challenge that is designing objective metrics that align with human perception.

State-of-the-art

Currently, several studies have evaluated the performance of a wide variety of FIQA algorithms, to the extent that there are now multiple ways to assess the same database and obtain similar results. Some algorithms use MOS as the ground truth, so the higher the MOS, the better the image quality evaluation. Others use DMOS (Difference Mean Opinion Score), where the lower the score, the better the image evaluation (Athar et al.).

These are three frameworks based on human opinion that demonstrate state-of-the-art performance, especially in the case of Full-Reference (FR) methods, as they achieve good correlation with human perception of quality. According to Athar et al., the best-performing FR algorithms include the Information Weighted Structural Similarity Index (IWSSIM), the Feature Similarity Index for Color Images (FSIMc), the Dissimilarity Structural Similarity (DSS), and the Visual Saliency-induced Index (VSI). For fusion-based classification aggregation methods, the Rank Aggregation of Scores (RAS) objective metric outperforms other fusion methods and some FR methods. It is particularly promising since it requires no training and offers robust performance across datasets.

In the case of No-Reference (NR) methods, the Codebook Representation for No-Reference Image Quality Assessment (CORNIA), the Higher-Order Statistics Aggregation (HOSA), and the Deep Image Quality (dipIQ) performed the best. However, in terms of perceptual quality prediction, precision, and computational complexity, these methods still fall significantly short of the performance achieved by FR methods.

According to T. Schlett et al., the highest-performing algorithms employ methodologies involving deep learning based on FR-integration, FR-based inference, FR-based ground truth training, and utility-agnostic training—the latter combined with explicit method fusion. Athar et al. confirm these findings and further emphasize the

effectiveness of classification aggregation fusion methods, even without prior training, when combined with FR.

Methods such as Peak Signal-to-Noise-Ratio (PSNR) and Structural Similarity Index Measure (SSIM) fail to capture perceptual distortions (Akter et al.). It can thus be concluded that IQA models based on model fusion significantly approach MOS values, with random forest-based fusion surpassing linear and PCA-based methods (Wang et al.). Consequently, new deep learning approaches, such as Learned Perceptual Image Patch Similarity (LPIPS) and Deep Image Structure and Texture Similarity (DISTS), have been proposed to better align with subjective evaluations. More specifically, by integrating metrics for various characteristics through machine learning, we can achieve a more robust and human-aligned image quality evaluation (T. Schlett et al.). To improve objective FIQA, numerous methods have been proposed—most of them based on fusion approaches. These methods combine the strengths of multiple metrics, and by applying techniques such as random forest or regression-based fusion, they produce a more accurate and reliable objective quality metric. While PSNR may capture signal degradation better, SSIM focuses on structural similarity, and LPIPS incorporates deep features aligned with perceptual judgements. By combining these, fusion methods have a more comprehensive view of image quality.

Some of the distortions that are used in this project are usually tested, such as JPEG compression, Gaussian blur and Motion blur (Wang et al.).

Several metrics are commonly used to evaluate image quality, each offering a different perspective. PSNR (Peak Signal-to-Noise Ratio) is a traditional metric that quantifies the ratio between the maximum possible signal and the noise in an image. While easy to compute, it often fails to align with human visual perception. SSIM (Structural Similarity Index Measure) addresses this limitation by evaluating luminance, contrast, and structural information between images, producing results that better reflect perceived image quality (Wang et al.). LPIPS (Learned Perceptual Image Patch Similarity) goes further by using deep neural networks to compare feature representations of images, offering a perceptual similarity measure that closely matches human judgment. Lastly, FID (Fréchet Inception Distance) assesses the quality of images, especially in generative tasks, by comparing the statistical distribution of features extracted from real and generated images using the Inception network. Unlike local similarity metrics, FID evaluates global similarity across image sets. Understanding the differences and

applications of these metrics is essential for choosing the most appropriate one for image quality assessment. (Atker et al.)

Although previous studies show strong performance for FR methods and metric fusion, few directly compare these to subjective MOS ratings in a demographically diverse dataset, which this work aims to do.

Even though existing literature shows promising results for FR and deep learning-based methods, few have validated these techniques using directly human opinion data. Most of them use older datasets which makes the results limited to real-world context.

Methodology

Dataset

For this study we used the publicly available Face Research Lab London (FRLL) Set, an ICAO compliant dataset that is composed by 102 neutral frontal facial images alongside metadata on attractiveness provided by over 2500 observers. It also has controlled acquisition conditions and availability of human attractiveness ratings.

We selected 10 images that were used as reference and on which we applied the seven different distortions, each applied at three different intensity levels, leading to a total of 210 distorted images. These are illustrated in Figure 1. The selected images were carefully chosen to have a wide and balanced range of ages, gender, ethnicities and attractiveness.

The distortions applied were:

- Motion Blur with vertical orientation with three levels (5, 15 and 25)
- Fisheye with three levels (0.15, 0.20 and 0.40)
- Facial Warp with three levels (1.3, 1.4 and 1.5)
- Gaussian Blur with three levels (0.2, 0.5 and 0.7)
- JPEG Compression with three levels (5, 15 and 30)
- Pincushion distortion with three levels (0.10, 0.20 and 0.25)
- Radial Distortion with three levels (0.4, 0.6 and 0.9

Figure 1- Examples of all distortions applied at different levels in the order described above. The top row shows the original images, while the bottom row shows the corresponding distorted images. The columns follow the same order of distortions as presented in the list.

The evaluations on each image were of, approximately, 25 per image, providing a reliable Mean Opinion Score dataset, as per instruction of ITU-R.

The assessment was made using a Single Stimulus test. Single stimulus is a type of **assessment** that consists in an evaluation of one image at a time. It is important not to compare quality between images, so that the perception can be the more realistic possible. In this test, the subject evaluates the **visual quality** of a series of individual images, assigning a **score** based on its **perceived quality** using a predefined scale.

The sessions of evaluation were conducted through a webapp, seen in Figure 2, where the subjects introduced both their demographic data (age, ethnicity, gender, etc.) and their non-demographic information (education level, device being used, and place where the test was being performed) and were asked to evaluate an image on a scale from 1 to 100, for as long as they wanted.

The webapp used was developed by a colleague of the ISR, André Neto.

In this test we reached a total of 4500 images evaluated, leading to an average of 25 opinions per image.

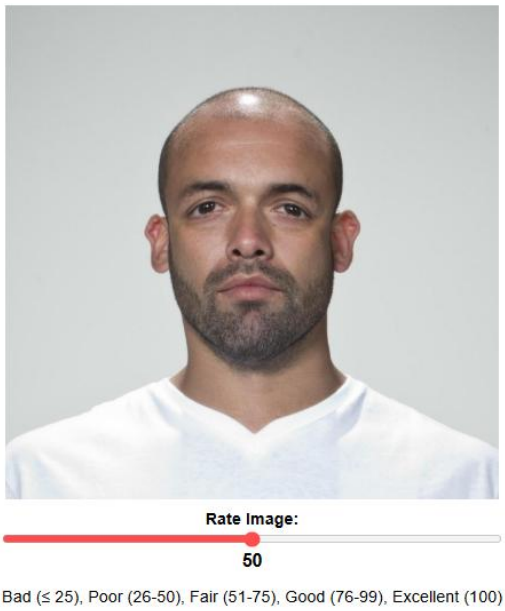

Figure 2 - Webapp used to assess image quality

Data Analysis

Image Quality Metrics

41 different Objective Image Quality Assessment metrics were evaluated, for example, Peak Signal-to-Noise Ratio (PSNR), SSIM, more traditional signal-based metrics, or LPIPS, that is a deep-learning-based approach.

To analyse the correlation between these metrics and MOS we used Pearson's and Spearman's correlation coefficients.

The Pearson correlation coefficient (PLCC) measure the strength and direction of a linear relationship between two variables. It ranges from -1 to +1, where -1 indicates a perfect negative linear relationship, 0 means no linear correlation and +1 means a perfect positive linear relationship. Pearson assumes that the data is normally distributed, that the relationship is linear and that there are no significant outliers.

The Spearman rank correlation coefficient (SRCC) assesses how well the relationship between two variables can be described using a monotonic function. Spearman works using ranked values. It also ranges from -1 to +1, where -1 represents a perfect reversal in ranks, 0 means no correlation and +1 perfect agreement in the ranking values. Spearman does not require the data to be normally distributed and is more robust to outliers.

Figure 3 presents the scatter plots that compare MOS values to individual IQA metrics, illustrating the relationship between subjective evaluations and objective IQA scores.

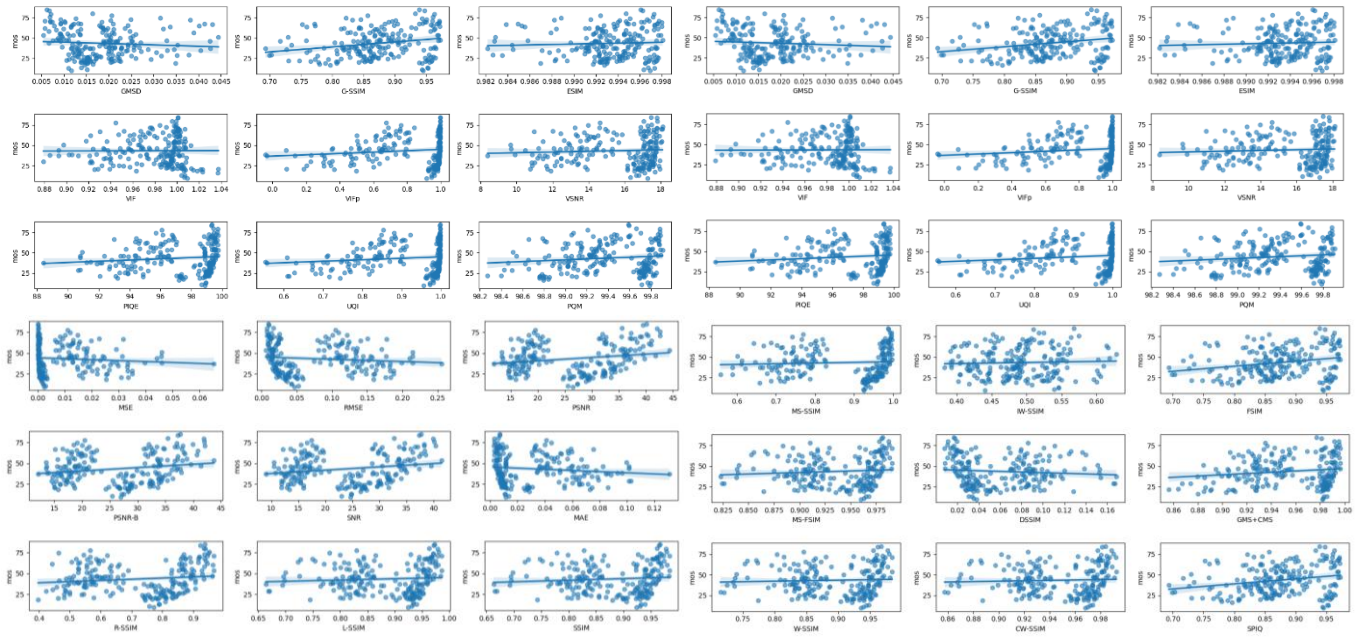

Figure 3 - MOS vs each objective metric

Pre-processing

Before applying the regressions on the data, we did a pre-processing, in which we eliminated outliers using the expression:

$$xi \notin [Q1 - 1{,}5 \cdot IQR, Q3 + 1{,}5 \cdot IQR] \Rightarrow xi\ e'\ outlier$$

Where:

- xi is an individual data point
- $Q1$ is the first quantile (25^{th} percentile) of the data
- $Q3$ is the third quantile (75^{th} percentile) of the data
- $IQR = Q3 - Q1$ is the interquartile range

We also applied a standardization using a class of the *scikit-learn* library on Python named *MinMaxScaler*, in which the minimal value was -1 and the maximum value was 1. This class was used since it provides a linear transformation and preserves the relative distribution of the data.

Fusion-based Method

Because the objective metrics alone don't perform well, we need to have a fusion-based approach to optimize MOS predictability.

To select which of the metrics were best to use in the fusion-based method, we used the Pearson Linear Correlation Coefficient (PLCC), the Spearman Rank Correlation Coefficient (SRCC), and linear regression tests. We chose the metrics that achieved the best scores across all three criteria. The results of this comparison are shown in Figure 4,

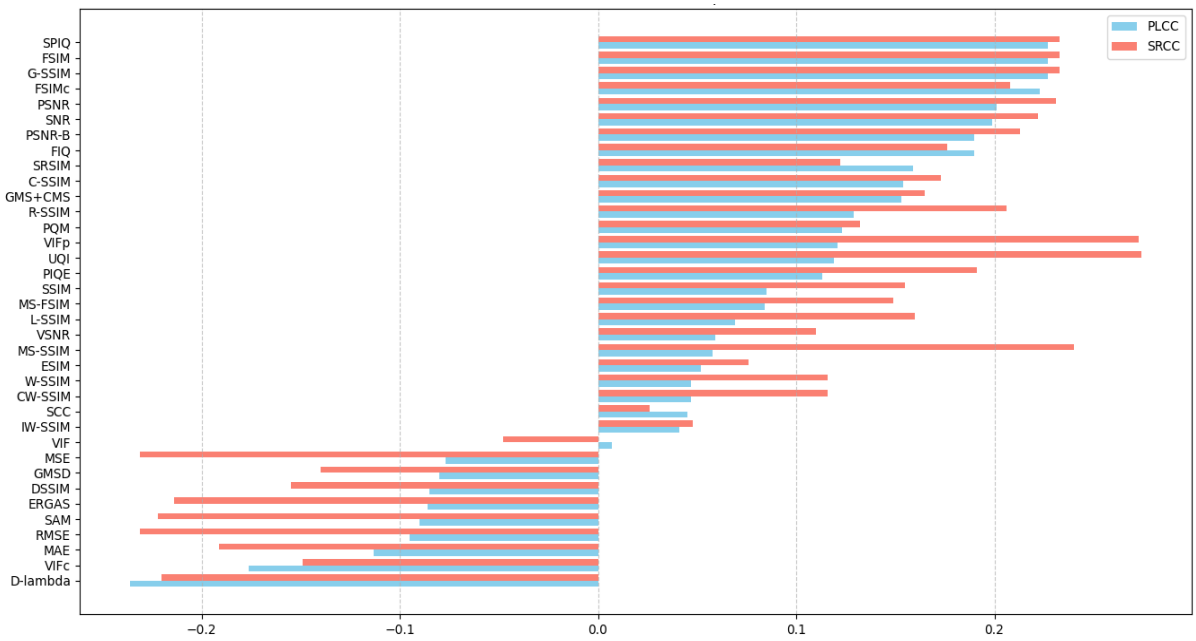

Figure 4- PLCC vs SRCC comparison for each objective metric

where PLCC and SRCC values are plotted on the horizontal axis. With that, we selected 10 metrics: FSIM, G-SSIM, SPIQ, PSNR, SNR, FSIMc, PSNR-B, FIQ, C-SSIM and SRSIM.

We did the fusion method with two different approaches: the random forest regression and the pseudoinverse of Moore Penrose.

Before starting implementing the two approaches we analysed the correlation between the 10 metrics previously selected through PLCC and SRCC tests. This led to a removal of 3 metrics which had strong correlation with each other. With that we continued with FSIM, PSNR, SNR, PSNR-B, FIQ, C-SSIM and SRSIM, that had a good correlation.

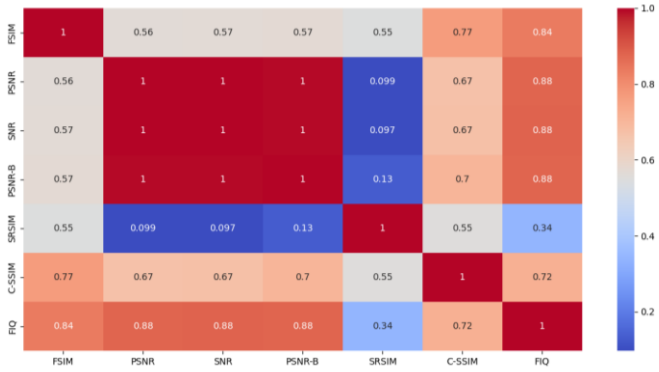

Figure 5 - Correlation between the 7 metrics selected to the Fusion models

Random Forest Regression

Random Forest Regression is a machine learning algorithm that is used to predict continuous values. It builds a forest of decision trees during training and then averages their outputs to make a final prediction. Every tree is trained on a random subset of the total data and uses a random subset of features, which leads to a reduction in overfitting.

The concept of this method is that, instead of relying on a single decision tree, a forest of many trees can give a more robust and accurate result. The randomness in the training process makes each tree different, and by averaging their predictions, the models benefit from their combined knowledge, cancelling their individual errors.

In order to improve the results of the random forest regression we had to apply a 10-fold cross validation. This consists in splitting the dataset in 10 equal parts, called folds. Then the model is trained on 9 folds and tested on the remaining fold. This process

is repeated 10 times, each time using a different fold for testing and the others for training. By doing this, we get a more reliable estimate of the model's performance and reduce the risk of overfitting.

The actual implementation of the Random Forest regressor was using the class *RandomForestRegressor* of the *sklearn.ensemble* library on *Python*.

After that, we removed data points with high residuals, meaning those where the difference between the actual values and the predictions from the random forest regressor was large. We then plotted the filtered dataset (with these outliers removed) and calculated the R^2 score to evaluate the model's performance on this cleaner data.

Pseudoinverse of Moore Penrose

The Moore-Penrose pseudoinverse is a generalization of the matrix inverse. It is used in cases where a matrix doesn't have an ordinary inverse, such as when it is not square or is singular (i.e. its determinant is zero). It allows us to still solve systems of equations or find best-fit solutions in cases where a regular inverse is undefined.

Specifically, for regression, the pseudoinverse is used to find a set of coefficients that best map input features to the target values. This is helpful in cases where the goal is to minimize the difference between the predicted and real values. The resulting regression equation is:

$$w = X^{+}.y = (X^{T}X)^{+}X^{T}y,$$

where X is the design matrix of input features, y is the vector of target values (e.g., MOS), and w is the resulting vector of regression coefficients. The plus signal indicates the Moore-Penrose pseudoinverse of the matrix.

To ensure robust evaluation, we used 5-fold cross-validation to reliably assess the regression model's performance and minimize overfitting.

The regression was performed using a linear regression model with polynomial features, fitted via the pseudoinverse method. The input features were first normalized to the range $[-1,1]$ and expanded to include polynomial terms of degree 2 to capture nonlinear relationships.

The implementation of the pseudoinverse was made using the function *np.linalg.pinv* of the NumPy library that uses the Singular Value Decomposition method to calculate the Moore Penrose pseudo-inverse. The formula via SVD is

$$A^{+} = V\Sigma^{+}V^{T}$$

Where U is an orthogonal matrix, Σ is a diagonal matrix with non-negative numbers (singular values) and V^{T} is the transpose of another orthogonal matrix. The + symbol indicated the pseudo-inverse of the matrix.

The coefficients of the model are then found by multiplying the values from the matrix by the vector for each image.

https://numpy.org/doc/stable/reference/generated/numpy.linalg.pinv.html

After training and evaluating the model through cross-validation, the model was retrained on the entire dataset to generate predictions for all samples. We then calculated the residuals (i.e., differences between predicted and true values) and removed the data that had the highest residuals.

Finally, we plotted the predicted versus true values, including the ideal y=x line and a linear regression line fitted on the filtered data. We also calculated the R^2 score on the filtered dataset to evaluate the model's performance after outlier removal.

Implementation Details

All the experiment were conducted using Python. The dataset went through a post-processing process to normalize IQA scores and eliminate outliers, to improve data quality and model reliability.

During the processing we used *MinMaxScaler* from the scikit-learn library to normalize the features within the range [-1,1]. We chose this method to preserve relative feature distances and maintain comparability between metrics with different scales.

To identify and remove samples with abnormally high residuals, we computed the z-score from the library *SciPy* of the residuals between predicted and actual Mean Opinion Scores (MOS). Observations with absolute z-score values greater than 2.5 were considered outliers and excluded. This filtering step was important to reduce the influence of extreme prediction errors that could bias model evaluation and distort regression fits.

To ensure a robust evaluation and minimize overfitting, we used a ten-fold cross validation for the random forest approach and a five-fold cross validation for the pseudoinverse for the training and evaluation. These techniques allowed us to assess the general performance of each method across multiple training and testing splits.

Results

Limitations of existing Objective Quality Metrics

In Figure 3, we see the different objective image quality metrics that were evaluated for their correlation with subjective Mean Opinion Scores (MOS). The individual objective metrics were later compared against MOS using Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Correlation Coefficient (SRCC) as shown in Figure 4.

More traditional metrics like PSNR and SSIM showed moderate correlation with MOS. Deep learning-based metrics like LPIPS demonstrated improved performance in approximating subjective evaluation but still exhibited inconsistencies across distortion types. This can be seen in Figure 2, since the same objective score corresponds to a wide range of MOS values indicating that the metric does not consistently reflect human perception of distortion severity. The scatter plots show that, despite some linear trends, individual metrics do not fully capture the variability observed in subjective assessments.

Correlation between fusion scores and MOS

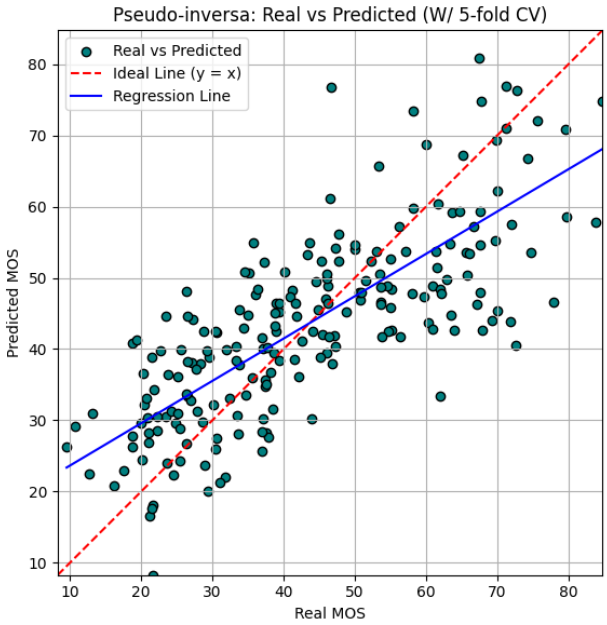

Figure 6 - Pseudo-Inverse with 5-fold validation compared to the ideal line (y=x) after pre-processing

The pseudoinverse regression model was trained with polynomial features, to fit not just linear combinations of the metrics, in order to better predict MOS. This model was evaluated using a 5-fold cross-validation. The model achieved and average cross-validation R^2 score of $0{,}287 \pm 0{,}103$, indicating a moderate ability to predict subjective scores from the fusion of objective metrics. After removing outlier with high residuals (Figure 6), the performance improved significantly. However, the model still showed limitations in general, with the regression line making an angle of 14,6 degrees from the ideal line. This suggest that the pseudoinverse regression is not the best approach to model the human perception of facial image quality, despite showing good results for this dataset.

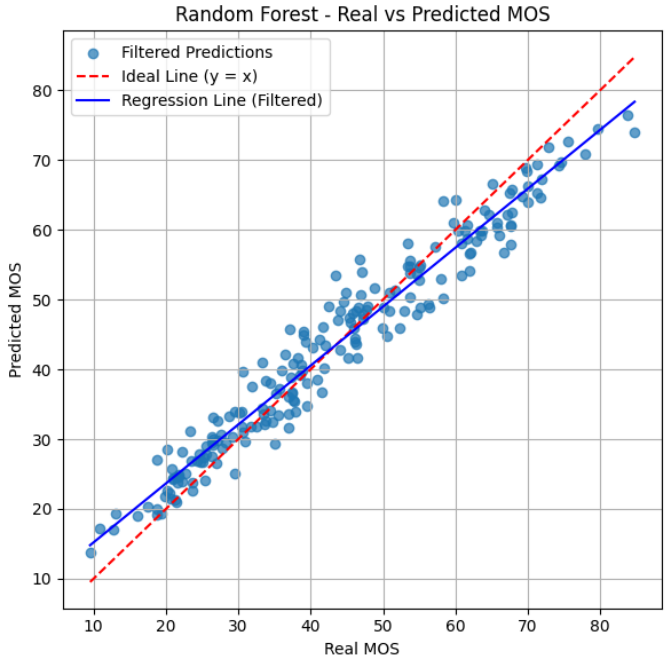

Figure 7 - Random Forest regression with a 10-fold cross validation after pre-processing

In contrast, the Random Forest Regression model, evaluated using 10-fold cross-validation, showed substantially better performance. It achieved and average R^2 score of $0{,}516 \pm 0{,}136$, demonstrating a stronger correlation with subjective MOS. After filtering out the high residuals data points, through the pre-processment previously mentioned, the performance of the model improved, reaching a final R^2 of 0,940, highlighting the model's robustness and an angle of 4,8 degrees with the ideal line.

This demonstrates that this approach is more accurate in approximating MOS than the pseudo-inverse, since it has a better alignment with the ideal line of y=x, indicating high agreement between predicted and real MOS values.

Comparison between fusion models

Figure 8 present a comparative visualization of the two fusion models. The Random Forest model clearly outperformed the pseudoinverse model, when comparing the two degrees with the ideal line, 14,6º for the pseudoinverse and 4,8º for the random forest. The Random Forest's multiple decision tress and random feature subsets contributed significantly to reducing overfitting and improving predictive performance. The fusion approach, regardless of the method improved substantially the performance of the model, performing substantially better than the individual IQA metrics.

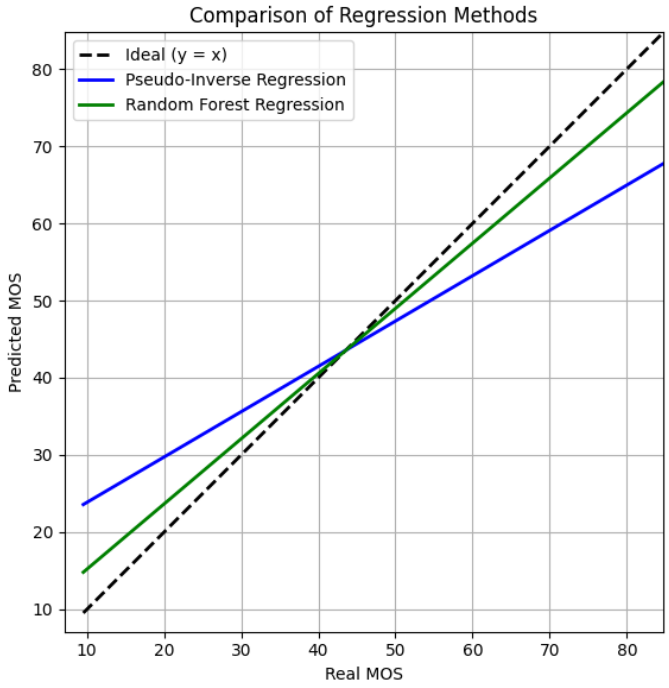

Figure 8 - Comparison between the two fusion models - Random Forest and Pseudo-Inverse of Moore Penrose

Discussion

The dataset used and the application of seven distinct distortion types at multiple intensities gave this study a wide and comprehensive study. We chose these images specifically to better reflect the demographic diversity of the real world, both race, gender and age. This was important because human perception of quality is known to be influenced by such variables. This approach ensures that the results were not biased by the monotony of one single variable, enhancing the robustness of the dataset.

The goal of this study was to explore the differences between human assessment of facial images and objective image quality metrics. We also wanted to evaluate whether combining these metrics could be a better way to approximate subjective judgment.

While the use of a single stimulus subjective evaluation method avoids direct comparison bias, it may still be affected by contextual factors, such as the user's environment or type of device.

As seen in the results, the more traditional metrics, though widely used, did not correlate strongly with the Mean Opinion Score (MOS). This aligns with previous studies that defend that these metrics are insufficient for perceptual tasks, especially with face images highly sensitive to distortions. (Athar et all.)

More advanced metrics, like LPIPS, which use deep neural networks, show comparatively better performance than the previous. The scatter plots presented confirmed that a single objective metrics cannot reliably distinguish high- and low-quality images as perceived by humans, what was confirmed by the weak alignment with the ideal y=x line. This indicates a limitation in the state-of-the-art models, since most of the IQA methods are designed to evaluate general images and aren't able to capture the perceptual quality of facial images.

To address these limitations, we applied fusion-based approaches. The goal was to use multiple metrics into a single predictive model. Two models were studied: a pseudoinverse regression model and a Random Forest regression model.

The pseudoinverse model, while useful for linear relationships, showed limited ability to model complex perceptual trends, achieving a relatively low to moderate R^2. This suggests that even with polynomial features, it struggles to capture the non-linear part of the data. This is in agreement with the literature, that indicates that perceptual image quality requires more flexible modelling.

In contrast, the random forest model achieved a very high R^2, after filtering high-residual data points, showing excellent performance in predicting MOS from objective features. This model's success can be attributed to its ensemble nature, i.e. the averaging multiple decision trees, which allows it to capture all the complex relationships without overfitting. Moreover, the use of the fold cross validation ensured robustness of the results.

Athar et al. and Wang et al. suggested that fusion-based methods, particularly ensemble-based models, tend to outperform single-metric evaluations, especially under full-reference conditions. Our results reinforce that claim, with the Random Forest model achieving an R^2 of 0.94, higher than previously reported on similar datasets using linear regressions alone.

All the results reinforce the observation, within this dataset, that no single metric fully captured the variability of human perception of facial image quality, whereas combining multiple metrics in a fusion model provided a closer approximation. However, these findings are limited to the specific images and distortions tested here and may not generalize to other datasets or contexts. The fusion models are more effective because they are able to capture the strongest qualities of each metrics and account for diverse image features.

Conclusion and Future Work

This study investigated the relationship between objective facial image quality metrics and subjective human perception, through Mean Opinion Scores (MOS). The results in this study showed that no existing single objective metrics reliably aligns with subjective perception, despite commonly used due to their simplicity. Deep learning-based metrics, such as LPIPS demonstrated improved alignment with human assessment, but showed significant discrepancies.

To overcome these limitations, we explored fusion metric approaches, comparing the Moore Penrose pseudoinverse regression with random forest regression model. The pseudoinverse model, despite capturing some linear trends, failed to model the full complexity of perceptual quality, achieving a modest R^2. In contrast, the Random Forest model reached a final R^2 of 0.94. This value indicates a strong agreement between the predicted and actual MOS, showing that this fusion approach is able to approximate human perception effectively.

These results are in agreement with the literature, suggesting that, in this context, combining multiple objective metrics in a fusion approach can approximate human perception more closely than individual metrics.

The main contributions of this study are the evaluation of objective IQA metrics on a facial dataset with demographic diversity, the validation of a robust Random Forest model, and the success in providing evidence that fusion methods can bridge the gap between objective metrics and subjective assessment.

Future work could focus on exploring deep learning-based models, including those that use neural regressors. Also expanding the distortions applied to the dataset, including a set of distortions closer to the real-world ones. Or even have more images in the dataset, with more diversity.

This approach can be integrated into biometrics acquisition systems to automatically reject low-quality samples and improving the performance of facial recognition technology.

Acknowledgments

I would like to thank and express my appreciation to my colleague André Neto for his continuous support throughout the project and for developing the webapp used to collect the opinions that formed the basis of this dataset. His help was essential to the success of the study.

References

S. Athar and Z. Wang, "A comprehensive performance evaluation of image quality assessment algorithms," IEEE Access, vol. 7, pp. 140030–140072, 2019.

T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch, "Face image quality assessment: A literature survey," ACM Computing Surveys, vol. 54, pp. 1–49, September 2022.

T. Liu, W. Lin, and C.-C. J. Kuo, "Image quality assessment using multi-method fusion," IEEE Transactions on Image Processing, vol. 22, no. 5, pp. 1793–1807, 2013.

H.-I. Kim, S. H. Lee, and Y. M. Ro, "Face image assessment learned with objective and relative face image qualities for improved face recognition," in Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 4027–4031, 2015.

D. Y. Tsao and M. S. Livingstone, "Mechanisms of face perception," Annual Review of Neuroscience, vol. 31, no. 1, pp. 411–437, 2008.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.

Z. Wang and Q. Li, "Image information and visual quality," IEEE Transactions on Image Processing, vol. 15, no. 2, pp. 430–444, 2006.

Y. Li, Y. Jiang, Z. Huang, X. Bai, L. Van Gool, and R. Timofte, "Blindsr: High-quality blind image super-resolution with pure vision transformer," arXiv preprint arXiv:2211.15265, 2022.P. Terhorst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness," March 2020.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” IEEE transactions on image processing, vol. 13, no. 4, pp. 600–612, 2004.

U. Sara, M. Akter, and M. Uddin, “Image quality assessment through fsim, ssim, mse and psnr—a comparative study,” Journal of Computer and Communications, vol. 7, no. 3, pp. 8–18, 2019.

J.-C. Yoo and C. Ahn, “Image matching using peak signal-to-noise ratio-based occlusion detection,” IET image processing, vol. 6, no. 5, pp. 483–495, 2012.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 586–595, 2018.

Neto, André and Gonçalves, Nuno. Pseudo-MOS Learning: A HybridFull-to-No-Reference FIQA Framework. 12th Iberian Conference on Pattern Recognition and Image Analysis, July 2025.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” 2018.

L. DeBruine and B. Jones, “Face Research Lab London Set,” 5 2017.