On the accuracy of objective image and video quality models: New methodology for performance evaluation

Lukáš Krasula, Karel Fliegel, Patrick Le Callet, Miloš Klíma

On the accuracy of objective image and video quality models: New methodology for performance evaluation

Lukáš Krasula*†, Karel Fliegel†, Patrick Le Callet*, and Miloš Klíma†
*LUNAM University - IRCCyN CNRS UMR 6597, 44306, Nantes, France
Email: {lukas.krasula, patrick.lecallet}@univ-nantes.fr
†Czech Technical University in Prague, Technická 2, 166 27 Prague 6, Czech Republic
Email: {krasuluk, fliegek, klima}@fel.cvut.cz

***Abstract*—There are several standard methods for evaluating the performance of models for objective quality assessment with respect to results of subjective tests. However, all of them suffer from one or more of the following drawbacks: They do not consider the uncertainty in the subjective scores, requiring the models to make certain decision where the correct behavior is not known. They are vulnerable to the quality range of the stimuli in the experiments. In order to compare the models, they require a mapping of predicted values to the subjective scores, thus not comparing the models exactly as they are used in the real scenarios. In this paper, new methodology for objective models performance evaluation is proposed. The method is based on determining the classification abilities of the models considering two scenarios inspired by the real applications. It does not suffer from the previously stated drawbacks and enables to easily evaluate the performance on the data from multiple subjective experiments. Moreover, techniques to determine statistical significance of the performance differences are suggested. The proposed framework is tested on several selected metrics and datasets, showing the ability to provide a complementary information about the models' behavior while being in parallel with other state-of-the-art methods.**

I. Introduction

The purpose of objective quality models is to substitute time consuming, expensive, and, in certain applications, impractical subjective quality tests. In order to determine the reliability of these models, their performances have to be evaluated with respect to the ground-truth data. The standard performance evaluation metrics considering experiments results in the form of Mean Opinion Scores (MOS) are described in ITU-T Rec. P.1401 [1].

The recommended methodology includes measuring linearity using Pearson Linear Correlation Coefficient (PLCC), prediction accuracy using Root Mean Squared Error (RMSE) and epsilon-insensitive RMSE (RMSE*), and Outlier Ratio (OR). Most of the studies also add Kendall Rank Order Correlation Coefficient (KROCC) and Spearman Rank Order Correlation Coefficient (SROCC) to determine the monotonicity of the predictions.

The main issue of the above stated methods (with the exception of RMSE*) is that they ignore the uncertainty of the subjective scores. Therefore they sometimes require the models to behave in a certain way, even though the subjective data are not statistically significant and we do not know what the correct behavior is. The methods are also designed for the cases where the range of quality levels considered within the subjective test is broad. If the quality range is narrow, the reliability can be significantly reduced due to the *range effect*, as described in [2].

Table I
RP measure for CSIQ database with 3rd order polynomial mapping using two different coefficients optimizations.

RP	**SSIM**	**IW-SSIM**
Coefficients optimized with RMSE	0.4164	0.3604
Coefficients optimized with PLCC	0.3804	0.3963

To overcome this, the measures of Resolving Power (RP) have been proposed in ITU-T Rec. J.149 [3]. The first measure finds the difference in predicted scores for stimuli A and B necessary to have 95% probability that the stimulus B is qualitatively better than the stimulus A. The model with lowest threshold is considered to be the most accurate. However, no information about the actual reliability of classification is provided. For that, classification plots are needed. They enable to graphically compare the behavior of correct classification with growing difference in predicted scores. Nevertheless, such comparison is not very practical for higher number of models being compared. Also, no method to determine a statistical significance of the results is defined.

Another significant drawback of performance evaluation measures (except for SROCC and KROCC) is the requirement for the predictions to be mapped to the subjective scores. Even though the same monotonic regression is used for all the models, the mapping strongly influences the models' behavior. To demonstrated the impact of mapping, we calculated RP measure for two popular objective indexes – structural similarity index (SSIM) [4] and its information weighted version (IW-SSIM) [5] on CSIQ database [6]. We used the 3rd order polynomial mapping with two different coefficients optimization methods – according to RMSE (as provided with the implementation of RP [3]) and according to PLCC (which is recommended by VQEG [7]). The results are stated in the Table I. The order of the scores changed. If another type of mapping (e.g. logistic) would be used, the influence can be even larger.

The initial purpose of mapping is the compression of the subjective scores at the ends of the scale but since different

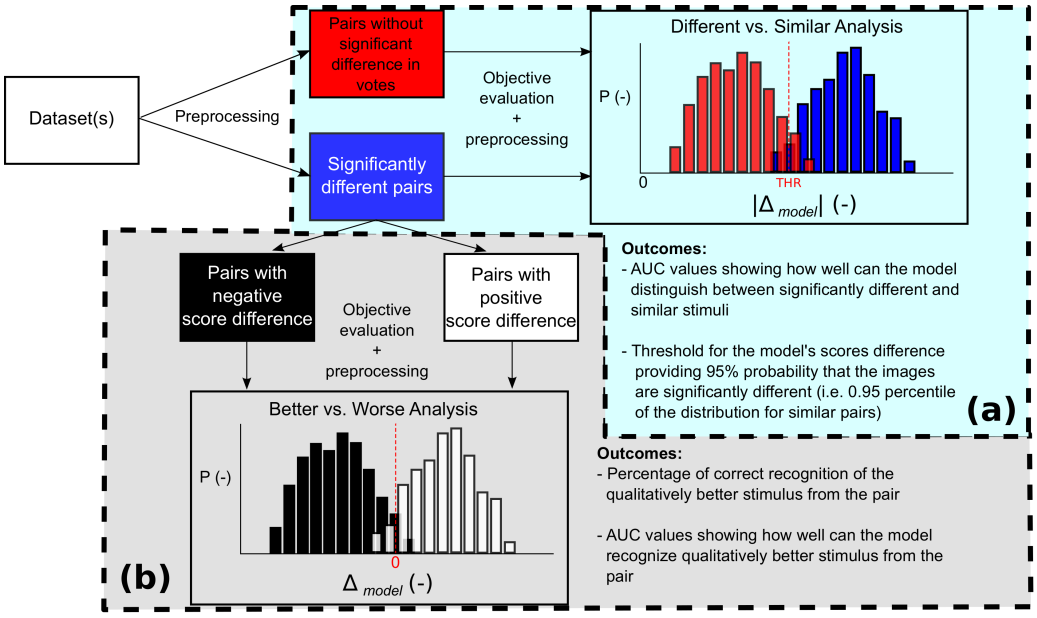


Figure 1. The framework of the proposed performance evaluation methodology.

functions can introduce such confusions, it appears to be more fair to find a method able to compare the models without mapping. Moreover, in real applications, the non-mapped scores are always used.

Considering the described issues, we propose a novel performance evaluation methodology that:

1) considers the statistical significance of the subjective scores;
2) is less dependent on the *range effect*;
3) compares the models as they are used in real applications without the necessity of mapping;
4) enables easy combination of data from different experiments;
5) provides the means to determine a statistical significance of differences in performance.

Section II of this paper describes the proposed method in detail, section III shows the analysis of the data from real experiments, and section IV concludes the paper.

II. Description of the Proposed Method

The basic assumptions of the presented method are similar to the RP measures [3]. The requirement for the objective models is to be able to reliably compare two stimuli and decide:

(a) *whether the stimuli are qualitatively different and*
(b) *if they are, which of them is of higher quality.*

We propose to evaluate the models' abilities considering the two above stated points separately, since some models could prove useful for one of the points but not the other. Certain scenarios then enable to use different models for individual tasks. For example, optimizing the bitrate while maintaining the perceived quality only requires the model to be reliable in the case (a). In enhancement, we want the final stimulus to be noticeably different (case (a)) and simultaneously of higher quality (case (b)) than the original. If we consider separate models for each case, conditional optimization can be used.

The whole framework of the proposed method is shown in Figure 1. The individual steps are described in the following subsections. Note that the method can also be used to benchmark the metrics from experiments not providing MOS-like results and the outcomes are not limited to the ones provided in Figure 1 [8]. In this paper, we report the most discriminative outcomes only. The implementation can be found on the author's webpage:
http://mmtg.fel.cvut.cz/personal/krasula/

A. Preprocessing of the Subjective Scores

The goal of the preprocessing stage is to determine what pairs in the dataset are statistically significantly different in the perceived quality. The ideal way to do this is to run ANOVA on the subjective scores and then use a post-hoc test, such as Tukey's honest significance difference (HSD) [9], to determine the statistically significant pairs. The advantage of this approach is that the post-hoc tests consider the problem of multiple comparisons.

However, most of the datasets only provide mean opinion scores (MOS) with respective standard deviations (SD) instead of the raw scores. In that case, similarly to [3], z-scores can be employed. We calculate a z-score for each pair of stimuli (i, j)

$$z(i,j) = \frac{|MOS(i) - MOS(j)|}{\sqrt{\frac{var(i)}{N(i)} + \frac{var(j)}{N(j)}}}, \quad (1)$$

where var is a variance of the votes and N is the number of observers who evaluated the given stimulus. The probability

that the stimuli are different is then calculated from the cumulative distribution function (cdf) of the normal distribution

$$p = \text{cdf}(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(\frac{z^2}{2}\right) dz. \quad (2)$$

We then consider the pairs with $p(i,j)$ higher than the selected significance level α to be significantly different. We will consider $\alpha = 0.95$ throughout this paper. The pairs are therefore divided into two groups – significantly different and similar.

The significantly different pairs are further divided into groups with positive and negative MOS difference. Since this is dependent on the order of stimuli in the pair, we recommend to consider each significantly different pair in both groups (reversed order) making the two groups symmetrical.

It can be seen, that we are taking only *binary* information about the stimuli pairs (different/similar and better/worse) from the subjective scores. This enables us to easily put data from different experiments together, regardless the method used for their obtaining, range, or format.

B. Preprocessing of the Predicted Scores

The objective models provide a predicted score for each stimulus in the dataset. We then obtain the differences of scores predicted by each $model$ for stimuli pair i and j as

$$\Delta_{model}(i,j) = score_{model}(i) - score_{model}(j), \quad (3)$$

where $score_{model}$ are the predicted values for the stimuli for particular $model$. Once the data has been preprocessed into the appropriate form, the performance evaluation according to the two cases described at the beginning of the section II can be executed.

C. Different vs. Similar Analysis

The first analysis is supposed to determine how well can the model distinguish between significantly different and similar pairs. The identification of these two groups of pairs has been described in section II-A.

The assumption is that the absolute difference of the predicted scores (i.e. their distance or L_2 norm) should be larger for the significantly different image pairs. The behavior of the well-performing model can look approximately like the example in the top of the Figure 1.

Typical method to determine the abilities of binary classifiers is Receiver Operating Characteristic (ROC) Analysis [10]. It creates a curve reflecting the correct classification when the threshold is shifted. The performance of the classifier can then be expressed as the Area Under the ROC Curve (AUC). Moreover, the threshold THR for the $|\Delta_{model}|$ leading to defined False Positive Rate (FPR), i.e. the probability that the pair is classified as different while being similar, can be determined. Note that these values depend on the range of values for the given model and therefore cannot be used for models' performance comparison. Nevertheless, they provide a valuable insight for the practical applications.

D. Better vs. Worse Analysis

The second analysis is performed on the significantly different pairs only. Here the goal is to determine whether the model is able to correctly recognize the stimulus of higher quality in the pair. The division of the significant image pairs into groups has been discussed in section II-A. An example of Δ_{model} values distributions for the two groups is shown in the bottom part of the Figure 1.

The most straightforward and determining factor is the percentage of correct classification in zero, showing how many times does the model correctly recognize the stimulus of higher quality. However, as will be shown in section III, the ROC analysis reflecting the overall behavior of the classification can provide some interesting insights as well.

E. Statistical Significance

When comparing multiple objective quality models, it is advisable to determine if the differences in performance are statistically significant. Literature provides several ways how to statistically compare AUC values from ROC analyses with different assumptions and power. In this paper, we use the method proposed by Hanley and McNeil [11].

The procedure is based on calculating a critical ratio c_{ab} between the AUC for models a and b. It is defined as

$$c_{ab} = \frac{AUC_a - AUC_b}{\sqrt{SE_a^2 + SE_b^2 - 2rSE_aSE_b}}, \quad (4)$$

where SE_a and SE_b are the standard errors for AUC_a and AUC_b, respectively, and r is an estimated correlation between the two areas given by the table in [11].

Standard error for each AUC can be computed according to [12] as

$$SE = \sqrt{\frac{AUC(1-AUC) + (n_{g1}-1)(Q_1 - AUC^2) + (n_{g2}-1)(Q_2 - AUC^2)}{n_{g1}n_{g2}}}, \quad (5)$$

where n_{g1} and n_{g2} are the numbers of elements in each group in the ROC analysis, and Q_1 and Q_2 are

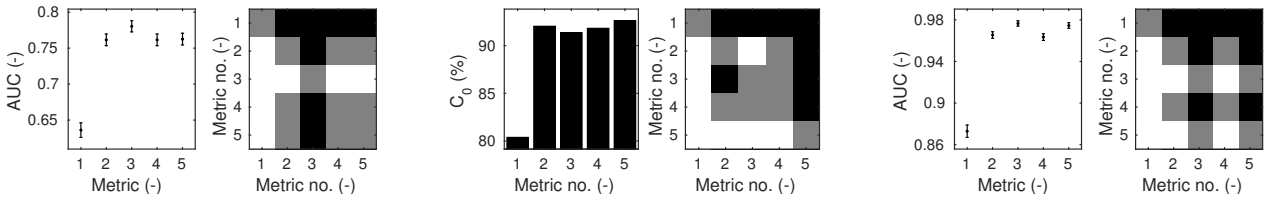
$$\begin{aligned} Q_1 &= AUC/(2 - AUC), \\ Q_2 &= 2AUC^2/(1 + AUC). \end{aligned} \quad (6)$$

The probability that the difference of the AUC_a and AUC_b is statistically significant can then be determined as $\text{cdf}(c_{ab})$ (see equation 2).

To statistically compare the percentage of correct recognition of the stimulus of higher quality (see section II-D), number of test can be used. Some possibilities are discussed in [8]. In our analyses, we employed Fisher's exact test [13].

If more than two models are being compared, the issue of *multiple comparisons* (i.e. Type I error propagation) should be considered. In this paper, we use a Benjamini-Hochberg procedure [14] to compensate for the error.

In the next Section, the analysis will be demonstrated on a real use case including several selected metrics and datasets.

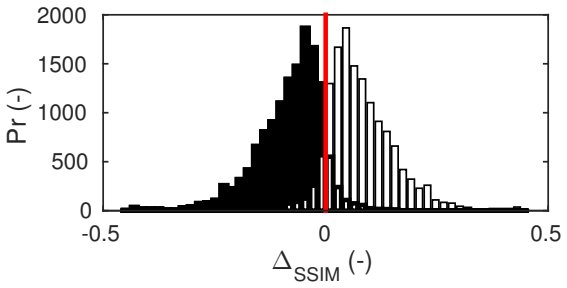


(a) Different vs. Similar (AUC) (b) Better vs. Worse (C_0) (c) Better vs. Worse (AUC)

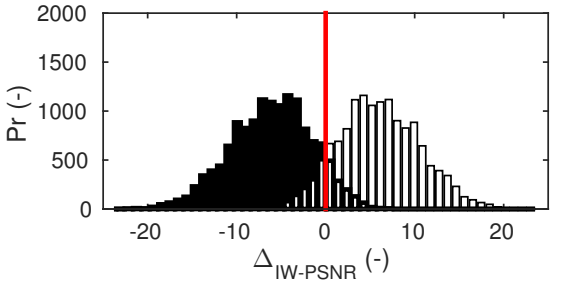
Figure 2. The results and statistical analysis for the IVC dataset. Significance plots show that the performance of the method in the row is either significantly better (white), lower (black), or none of the previous (gray).

Table II
NUMBERING OF THE OBJECTIVE METRICS.

1	2	3	4	5
PSNR	SSIM	IW-PSNR	MS-SSIM	IW-SSIM

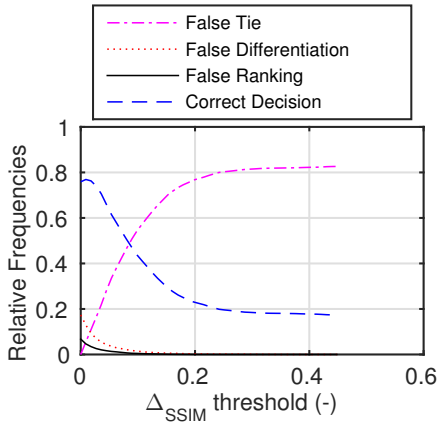


(a) SSIM

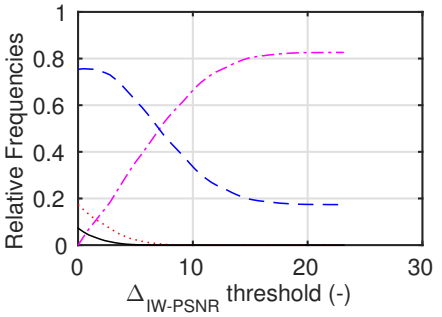


(b) IW-PSNR

Figure 3. The distributions for the two groups in Better vs. Worse Analysis.



(a) SSIM



(b) IW-PSNR

Figure 4. Classification Plots as defined in [3].

III. USE CASE STUDY

To demonstrate the use of the proposed methodology practically, we decided to work with the data available from the performance evaluation of Information Weighted SSIM (IW-SSIM) metric [5]. This way, no additional bias in data obtaining can be introduced and the outcomes of other performance evaluation methods are available for comparison. All the data were obtained from the supporting website[1].

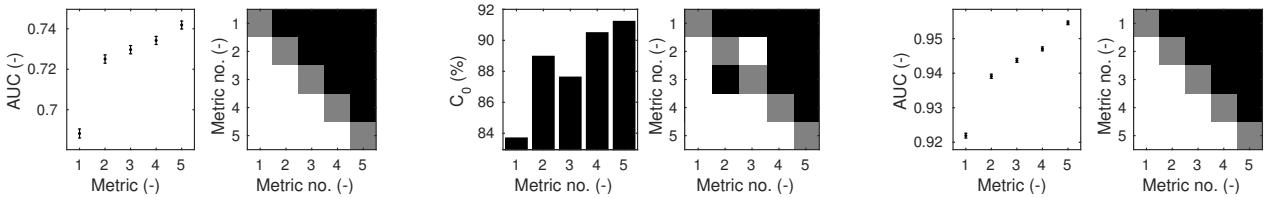
Predicted scores for five objective algorithms are provided – PSNR, SSIM [4], Information Weighted PSNR (IW-PSNR) [5], Multiscale SSIM (MS-SSIM) [15], and IW-SSIM [5]. In the following figures, the algorithms will be numbered according to the table II. The datasets used to evaluate their performance in [5] are LIVE [16], A57 [17], IVC [18], Toyama [19], TID2008 [20], and CSIQ [6].

For the detailed demonstration of the proposed approach, IVC dataset has been selected, since it exhibits an interesting behavior of the tested algorithms.

A. Performance on IVC dataset

The results of the particular analyses (sections II-C and II-D), namely AUCs and percentage of correct classification (C_0) with statistical significance of differences, are depicted in Figure 2. The error bars represent 95% confidence intervals. The white boxes in the significance plots correspond to the cases when model in the row significantly outperforms the model in the column. If its performance is significantly lower, the corresponding box is black. The gray box symbolizes the case where we are not able to determine the better performing method.

[1]https://ece.uwaterloo.ca/~z70wang/research/iwssim/



(a) Different vs. Similar (AUC) (b) Better vs. Worse (C_0) (c) Better vs. Worse (AUC)

Figure 5. The results and statistical analysis for the four datasets. Significance plots show that the performance of the method in the row is either significantly better (white), lower (black), or none of the previous (gray).

Table III
AUC VALUES FOR DIFFERENT VS. SIMILAR ANALYSIS

AUC	**PSNR**	**SSIM**	**IW-PSNR**	**MS-SSIM**	**IW-SSIM**
IVC	0.6360	0.7615	**0.7803**	0.7615	0.7626
CSIQ	0.6827	0.6910	0.7064	0.7056	**0.7148**
LIVE	0.7250	0.7654	**0.7969**	0.7625	0.7744
Toyama	0.5954	0.7068	0.7182	0.7117	**0.7563**
ALL	0.6882	0.7251	0.7297	0.7342	**0.7419**

Table IV
THRESHOLDS FOR 5% FPR OF CLASSIFYING PAIR AS DIFFERENT.

THR	**PSNR**	**SSIM**	**IW-PSNR**	**MS-SSIM**	**IW-SSIM**
IVC	8.4461	0.1285	6.4705	0.0757	0.1023
CSIQ	13.0470	0.2818	19.0379	0.2198	0.2686
LIVE	9.7317	0.2677	11.0294	0.1713	0.1710
Toyama	10.6431	0.0873	8.6840	0.0444	0.0429
ALL	12.4806	0.2677	17.9861	0.2002	0.2429

Table V
CORRECT CLASSIFICATION IN BETTER VS. WORSE ANALYSIS

C_0	**PSNR**	**SSIM**	**IW-PSNR**	**MS-SSIM**	**IW-SSIM**
IVC	0.8038	0.9203	0.9135	0.9181	**0.9261**
CSIQ	0.8279	0.8721	0.8542	0.8978	**0.9049**
LIVE	0.8518	0.9081	0.8998	0.9122	**0.9190**
Toyama	0.7630	0.9069	0.8841	0.9127	**0.9386**
ALL	0.8369	0.8897	0.8762	0.9049	**0.9122**

Table VI
AUC VALUES FOR BETTER VS. WORSE ANALYSIS

AUC	**PSNR**	**SSIM**	**IW-PSNR**	**MS-SSIM**	**IW-SSIM**
IVC	0.8877	0.9669	**0.9795**	0.9649	0.9768
CSIQ	0.9140	0.9227	0.9265	0.9357	**0.9444**
LIVE	0.9377	0.9568	0.9657	0.9597	**0.9660**
Toyama	0.8570	0.9657	0.9613	0.9685	**0.9843**
ALL	0.9219	0.9391	0.9437	0.9470	**0.9546**

It can be seen that IW-PSNR (#3) significantly outperforms all the other metrics in the first analysis (Figure 2(a)). On the other hand, PSNR (#1) has the lowest performance, also with statistical significance.

In the second analysis, we can observe an interesting phenomenon. IW-PSNR (#3) provides statistically worse classification than SSIM (#2) (Figure 2(b)) but reaches significantly higher AUC value (Figure 2(c)). To explain this, we closely studied the behavior of the two metrics. The histograms of Δ_{SSIM} and $\Delta_{\text{IW-PSNR}}$ for the two groups defined in section II-A are depicted in Figure 3 (the number of bins is the same for both models).

The distributions for the IW-PSNR are much broader with modes more distant from 0. This means that if we broaden the red area in the Figure 3, which is equivalent to not considering the pairs with small differences as being different, the performance of IW-PSNR will be dropping slower than in case of SSIM. This exactly reflects the information that can be extracted from the Classification Plots [3] (Figure 4).

The implementation of the Classification Plots was obtained directly from the recommendation and the version with predicted scores non-mapped to the subjective ones is used, thus not allowing for numerical comparison. Nevertheless, visual comparison confirms that the curve for correct classification is broader for IW-PSNR but reaching a lower maximal value. Our approach is therefore able to quantify this effect and allows for numerical comparison without the need for any mapping.

Moreover, the percentage of correct classification (C_0) agrees with other performance comparison techniques, such as SROCC, calculated in [5]. The proposed analyses therefore reach similar results as state-of-the-art methods while simultaneously providing more insight into the models' behavior.

B. Performance on Multiple Datasets Together

To evaluate the performance on data from multiple experiments, weighted average of the particular performance evaluation methods is usually taken. This is not possible for all the methods (e.g. values of RMSE and RMSE* depend on the range of subjective scores). Moreover, given the different uncertainties of the particular results, the final statistical comparison is not well defined.

In the case of the RP measures, the necessity of mapping the data to the common scale makes the different experiments combination impractical. Since we extract only the *binary* information from the subjective data, the combination becomes much easier because it represents simple adding more stimuli pairs in the groups (different/similar, better/worse). All analyses are therefore performed on all the data together.

In this section, we demonstrate the performance evaluation on four databases together (IVC [18], CSIQ [6], LIVE [16], and Toyama [19]). We do not use A57 and TID2008 datasets here, since the subjective data for the former are obtained from the seven expert subjects only, making the statistical processing not very relevant. The latter is omitted because it is not possible to determine how many observers evaluated each image from the description. Only overall number of observers

is provided but not all of them evaluated all the content. Computation of z-scores would therefore be unreliable.

The results for the four databases are depicted in Figure 5. The Tables III-VI contain the final values obtained from the datasets separately, as well as from their combination. Note that in Table IV, we also report the thresholds for the $|\Delta_{model}|$ necessary to ensure the 5% FPR (i.e. for 5% probability that the pair is classified as different while being similar). The values are dependent on the range of models' values and therefore cannot be directly used for models' comparison but their differences for particular datasets provide another insight and they are important for the practical use of the models.

Several conclusions can be drawn from the overall results. Firstly, the best performing model is IW-SSIM, followed by MS-SSIM. We can see in the Table III that even though the IW-PSNR metric reaches higher AUC value than MS-SSIM in the Different vs. Similar Analysis for each database separately, the overall performance of MS-SSIM is higher. This shows that weighted average of the particular results does not have to lead to the same conclusions as analysing all the data at the same time.

Also the effect described in the Section III-A where SSIM provides better classification in the Better vs. Worse Analysis but the AUC value is higher for IW-PSNR is reflected in the overall results as well. For the explanation, refer to the stated Section.

For most of the metrics, CSIQ database appears to be the most challenging. The only exception from this is PSNR which works the worst for Toyama dataset. These findings are in parallel with correlation measures from [5].

The last observation we will provide is the room for improvement in models' abilities with respect to the the Different vs. Similar Analysis. Although it is true that not all well-performing objective methods has been tested here. Nevertheless, IW-SSIM is considered to be one of the reliable models, outperforming other popular metrics [5], and the overall AUC value of 0.7419 is not very high.

IV. Conclusion

We presented a novel methodology for performance evaluation of objective models inspired by the real applications. The method does not require any mapping to enable numerical comparisons, takes into account statistical significance of subjective scores, depends less on the quality range of the dataset, enables easy combination of data from different subjective experiments, and provides means to determine statistical significance of the performance differences.

It has been demonstrated and analysed in detail on five objective models. It has been shown that the methodology provides a complementary information about the models' behavior while being in parallel with other state-of-the-art techniques and simultaneously enables for simple comparisons.

The ability to easily and meaningfully combine data from multiple experiments has been presented. The result suggests that the averaging of results (possibly weighted according to the dataset size) can lead to different conclusions than the analysis on all the data. Moreover, our methodology maintains the possibility to analyse the data statistically even after merging results from multiple experiments.

References

[1] ITU-T Recommendation P.1401, *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*, ITU-T Std., 2012.

[2] M. A. Saad, P. Le Callet, and P. Corriveau, "Blind image quality assessment: Unanswered questions and future directions in the light of consumers needs," *VQEG e-letter*, vol. 1, no. 2, pp. 62–66, December 2014. [Online]. Available: ftp://vqeg.its.bldrdoc.gov/eLetter/Issues/VQEG_eLetter_vol01_issue2.pdf

[3] ITU-T Reccomendation J.149, *Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM)*, J.149 Std., 2004.

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

[5] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185 – 1198, May 2011.

[6] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, p. 011006, 2010. [Online]. Available: http://link.aip.org/link/?JEI/19/011006/1

[7] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," VQEG, Tech. Rep., 2000.

[8] P. Hanhart, L. Krasula, P. Le Callet, and T. Ebrahimi, "How to benchmark objective quality metrics from paired comparison data," in *International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

[9] J. Tukey, "Comparing individual means in the analysis of variance," *Biometrics*, vol. 5, no. 2, pp. 99–114, 1949.

[10] J. A. Swets, *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers.*, J. A. Swets, Ed. Mahwah, New Jersey: Lawrence Erlbaum Associates, 1996.

[11] J. A. Hanley and B. J. McNeil, "A method of comparing the area under two ROC curves derived from the same cases," *Radiology*, vol. 148, pp. 839–843, 1983.

[12] ——, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29–36, 1982.

[13] R. A. Fisher, "On the interpretation of χ^2 from contingency tables, and the calculation of p," *Journal of Royal Statistical Society*, vol. 85, no. 1, pp. 87–94, 1922.

[14] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, vol. 57, no. 1, pp. 289–300, 1995.

[15] Z. Wang, E. Simoncelli, and A. Bovik, "Multi-scale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signal, Systems and Computers*, vol. 2, 2003, pp. 1398–1402.

[16] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[17] D. M. Chandler and S. S. Hemami, "VSNR: A waveletbased visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284–2298, September 2007.

[18] A. Ninassi, P. Le Callet, and F. Autrusseau, "Pseudo no reference image quality metric using perceptual data hiding," in *SPIE Human Vision and Electronic Imaging*, vol. 6057, 2006, [online]. Available: http://www2.irccyn.ec-nantes.fr/ivcdb.

[19] Z. M. P. Sazzad, Y. Kawayoke, and Y. Horita. Image quality evaluation database. [Online]. Available: http://mict.eng.u-toyama.ac.jp/database_toyama/

[20] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "Tid2008 – a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009, [online]. Available: http://www.ponomarenko.info/tid2008.htm.