Algorithm Selection for Image Quality Assessment

Markus Wagner
University of Adelaide
Adelaide, Australia
markus.wagner@adelaide.edu.au

Hanhe Lin
University of Konstanz
Konstanz, Germany
hanhe.lin@uni-konstanz.de

Shujun Li
University of Kent
Canterbury, Kent, UK
S.J.Li@kent.ac.uk

Dietmar Saupe
University of Konstanz
Konstanz, Germany
dietmar.saupe@uni-konstanz.de

Abstract—Subjective perceptual image quality can be assessed in lab studies by human observers. Objective image quality assessment (IQA) refers to algorithms for estimation of the mean subjective quality ratings. Many such methods have been proposed, both for blind IQA in which no original reference image is available as well as for the full-reference case. We compared 8 state-of-the-art algorithms for blind IQA and showed that an oracle, able to predict the best performing method for any given input image, yields a hybrid method that could outperform even the best single existing method by a large margin. In this contribution we address the research question whether established methods to learn such an oracle can improve blind IQA. We applied AutoFolio, a state-of-the-art system that trains an algorithm selector to choose a well-performing algorithm for a given instance. We also trained deep neural networks to predict an algorithm selector to choose a well-performing algorithm for a given input image the best suited IQA method out of a portfolio of a set of candidate algorithms.

This is an instance of the general algorithm selection problem \( \Pi \). Given a portfolio \( \mathcal{P} \) of algorithms or methods, a set \( \mathcal{I} \) of problems, and a cost metric \( m : \mathcal{P} \times \mathcal{I} \rightarrow \mathbb{R} \), the algorithm selection problem consists of finding a mapping \( s : \mathcal{I} \rightarrow \mathcal{P} \) from instances in \( \mathcal{I} \) to algorithms in \( \mathcal{P} \) such that the total cost \( \sum_{I \in \mathcal{I}} m(s(I), I) \) across all instances is minimized. If \( \mathcal{I} \) and \( \mathcal{P} \) are finite, the single best method (SBM) is given by \( M^* \in \mathcal{P} \) with \( M^* = \arg \min_{M \in \mathcal{P}} \sum_{I \in \mathcal{I}} m(M, I) \), and the virtual best selection model (VBM), also called the oracle, \( O \), is the one that selects the best algorithm in each case, so \( O(I) = \arg \min_{M \in \mathcal{P}} m(M, I) \) for all \( I \in \mathcal{I} \).

II. DATA SET AND THE VIRTUAL BEST METHOD

In [4] the authors introduced a diverse dataset, called KonQ-10k, of 10,073 natural images with authentic distortions intended for machine learning BIQA methods. Currently, it is the largest such dataset available. It is subdivided into a training...
TABLE I
PERFORMANCE OF 8 IQA METHODS ON THE KONIQ-10K TEST SET.

| Method    | Features | SROCC | MAE | Method best for images |
|-----------|----------|-------|-----|------------------------|
| BIQI      | 18       | 0.559 | 8.339 | Rank 1: SSEQ, Rank 2: BRISQUE, Rank 3: CORNIA |
| BLIINDS-II| 24       | 0.585 | 9.239 |                          |
| BRISQUE   | 36       | 0.705 | 8.224 |                          |
| CORNIA    | 20,000   | 0.780 | 7.308 |                          |
| DIIVINE   | 88       | 0.589 | 8.180 |                          |
| HOSA      | 14,700   | 0.805 | 6.792 |                          |
| SSEQ      | 12       | 0.604 | 9.403 |                          |
| KonCept512| 1,536    | 0.921 | 4.154 |                          |

Virtual best method: NA 0.978 2.069 2,015 0 0

We have fitted the predictions of the eight methods to the ground truth values of the training set, which were scaled to the interval $[0, 100]$, by nonlinear regression, using the 5-parameter logistic function from [5]. This is a necessary preprocessing step before algorithm selection, because IQA methods generally are trained to give the best correlation with ground truth rather than minimizing an average error measure. In Table II we list the Spearman rank order correlation coefficient (SROCC) and the mean absolute error (MAE) of the predictions of all methods. The MAE is based on the joint quality scale $[0, 100]$.

After this alignment, we obtained the virtual best method by checking for each of the 2015 test images which method estimated its quality closest to the ground truth MOS. The columns labeled “Rank 1, 2, 3” in Table I show the numbers of images for which each method gave the best, the second and the third best result. The single best method, KonCept512, provided 682 out of 2,015 scores for the virtual best method. This is more than three times as many as any other method, but still only 33.8% of all test images. The virtual best method gave an SROCC value of 0.978 and an MAE of 2.069, much better than the single best method. The correlation diagram and scatter plots in Figure I show a certain degree of complementarity of the algorithms, which, in principle, should allow us to train an effective algorithm selector.

III. ALGORITHM SELECTION USING AUTOFOLIO

In our first attempt of algorithm selection for BIQA, we employed AutoFolio [2]. This tool automatically determines a well-performing algorithm selection approach and its hyperparameters. In its learning phase, AutoFolio takes as input two matrices: one that lists for each training instance its instance feature values, and the other one lists for each instance the performance of all (eight) algorithms. AutoFolio takes these and then explores the “algorithm selector design space”, which includes design parameters such as different models (e.g., random forests and XGBoost) with various parameterizations, and preprocessing options (e.g., PCA on/off and scaling on/off).

From Table II the total number of features is 36,414, mostly because CORNIA, HOSA, and KonCept512 make use of many features. To limit a possible selection bias of features by AutoFolio and to reduce complexity, we performed a principle component analysis for the set of features of each of the three methods mentioned above and then selected the most important 100 features in each case. In total, this resulted in 478 features that the eight methods contribute.

Table III lists the results of two experiments. In both, AutoFolio was given 24 hours to explore the model space. In the first one, we allowed it to use all eight algorithms. Despite our and AutoFolio’s best efforts (it explored over 500 models in 24 hours), the best algorithm selector chose KonCept512 for all of the 2015 test instances, even though the VBM would pick it in just about 34% of all cases. Due to KonCept512’s dominance, we excluded it in the second experiment. The MAE of the remaining seven method’s VBM increased to 3.063. Interestingly, AutoFolio now managed to learn an algorithm selector that performed slightly better than the single best method of the remaining seven algorithms. However, this holds only for the MAE performance metric.
V. DISCUSSION AND CONCLUSION

The virtual best algorithm by means of algorithm selection from a portfolio of eight methods would yield an extreme improvement of IQA performance over the single best one, KonCept512 (SROCC of 0.978 versus 0.921). However, all our attempts to apply methods of algorithm selection have failed to achieve a performance better than that of the single best one. Using state-of-the-art algorithm selection, the best model came out to be equal to the best single method, KonCept512. Moreover, both approaches to learning to identify the best IQA method for an input image by deep neural networks gave results on the test set that are worse than those of the single best method (SROCCs of 0.871 and 0.908).

Our explanation is a combination of two issues. Firstly, we conjecture that the performance of the single best algorithm, KonCept512, is already close to being optimal, i.e., at the saturation limit of what can be achieved for blind IQA on our training and test sets. Secondly, we conjecture that the clear superiority of the virtual best algorithm may be attributed to ‘noisy’ evaluation of image quality. Consider an IQA method and a fixed test image. For this image there are numerous other images that are perceptually indistinguishable but different in terms of pixel RGB values. When evaluating an IQA method on this set of visually equivalent images, we would obtain a distribution of image quality values. So the actual quality estimate of a particular image can be interpreted as the mean value of all of these measurements, plus an added noise term. In this case the virtual best method can still achieve an improvement over the optimal method, but only due to exploitation of noise which, of course, cannot be predicted by any machine learning on a training set.

Therefore, our work, although providing a negative answer to the initial question of whether algorithm selection can improve blind image quality assessment, opens up a number of interesting new research questions: Can one quantitatively and reliably assess the noisiness of IQA methods? Does denoising of IQA methods improve their performance? And finally, does denoising remove the large gap between the single best method and the virtual best, and are denoised IQA methods better suited for the algorithm selection strategy?

REFERENCES

[1] J. R. Rice, “The algorithm selection problem,” in Advances in Computers. Elsevier, 1976, vol. 15, pp. 65–118.
[2] L. Xu, W. Lin, and C.-C. J. Kuo, “Metrics fusion,” in Visual Quality Assessment by Machine Learning, ser. SpringerBriefs in Electrical and Computer Engineering. Springer Singapore, 2015, ch. 5, pp. 93–122.
[3] M. Oszust, “Decision fusion for image quality assessment using an optimization approach,” IEEE Signal Processing Letters, vol. 23, no. 1, pp. 65–69, 2016.
[4] H. Lin, V. Hosu, and D. Saupe, “KonIQ-10K: Towards an ecologically valid and large-scale IQA database,” arXiv:1803.08489 (cs.CV), 2018.
[5] 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.

[6] B. Bischl, P. Kerschke, L. Kotthoff, M. Lindauer, Y. Malitsky, A. Frechette, H. Hoos, F. Hutter, K. Leyton-Brown, K. Tierney, and J. Vanschoren, “Aslib: A benchmark library for algorithm selection,” *Artificial Intelligence Journal (AIJ)*, vol. 237, pp. 41–58, 2016.

[7] M. Lindauer, H. Hoos, F. Hutter, and T. Schaub, “Autofolio: An automatically configured algorithm selector,” *Journal of Artificial Intelligence Research*, vol. 53, pp. 745–778, 2015.

[8] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, Inception-ResNet and the impact of residual connections on learning,” in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 4, 2017, p. 12.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.