

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 the best method. Our results did not give a positive answer, algorithm selection did not yield a significant improvement over the single best method. Looking into the results in depth, we observed that the noise in images may have played a role in why our trained classifiers could not predict the oracle. This motivates the consideration of noisiness in IQA methods, a property that has so far not been observed and that opens up several interesting new research questions and applications.

Index Terms—image quality assessment, algorithm selection, machine learning, deep learning

I. INTRODUCTION

The perceptual quality of visual media is of relevance for the development of media compression and enhancement algorithms as well as for content providers wishing to ensure sufficient user satisfaction. Assessment of visual quality requires human judges or algorithmic (“objective”) methods. These can be trained on subjective mean opinion scores (MOS) from benchmarks achieved by human lab assessments, or, more recently, by larger crowdsourcing studies.

In this contribution we consider the case of blind image quality assessment (BIQA), i.e., the estimation of subjective perceptual image quality without availability of a pristine reference image. There are several image quality datasets available for training and testing BIQA methods, and a number of algorithms have been proposed to solve the BIQA task providing more or less accuracy. It can be expected that there is no single method that achieves the best result, i.e., a quality estimation nearest to the MOS, for all of the images in a test set. Therefore, here we consider learning to predict for each

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 [1]: 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, \mathcal{O} , is the one that selects the best algorithm in each case, so $\mathcal{O}(I) = \arg \min_{M \in \mathcal{P}} m(M, I)$ for all $I \in \mathcal{I}$.

In the case of BIQA, the set of algorithms is finite, $M_k \in \mathcal{P}, k = 1, \dots, K$, and the instances are images from a test set of a benchmark dataset, $I_n \in \mathcal{I}, n = 1, \dots, N$, where the image qualities have been assessed by mean opinion scores, $\text{MOS}(I_n)$. The cost function can be, e.g., the absolute error of the BIQA method, $s(I, M) = |M(I) - \text{MOS}(I)|$.

It turns out that for a set of state-of-the-art BIQA algorithms and a large-scale image quality dataset there is a very large performance gap between the single and the virtual best method. Thus, in this work we are pursuing the research question for BIQA, whether advanced methods of algorithm selection are able to close this gap. All our attempts, however, failed in this regard. It seems, algorithm selection for BIQA does not yield an improvement over the single best method. Although negative, this result gives rise to a number of interesting new research questions, posed at the end.

We are not aware of any previous work on algorithm selection for blind IQA as well as for the full reference case (FR-IQA). However, hybrid method have been proposed for FR-IQA, combining all methods from a portfolio by linear combinations trained by regression for the quality assessment. In addition, images can be classified according to distortion type and for each of these a separate method fusion can be designed [2]. It was also proposed for FR-IQA to select and linearly combine a subset of IQA methods, however, globally, i.e., not adaptively for each input image [3].

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

Presented at the Seventh Workshop on COnfiguration and SElection of ALgorithms (COSEAL), Potsdam, Germany, August 26–27, 2019.

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project-id 251654672 – TRR 161

TABLE I
PERFORMANCE OF 8 IQA METHODS ON THE KONIQ-10K TEST SET.

Method	Features	SROCC	MAE	Method best for images		
				Rank 1	Rank 2	Rank 3
BIQI	18	0.559	8.339	187	188	240
BLIINDS-II	24	0.585	9.239	185	215	205
BRISQUE	36	0.705	8.224	176	205	253
CORNIA	20,000	0.780	7.308	217	263	286
DIIVINE	88	0.589	8.180	169	198	259
HOSA	14,700	0.805	6.792	220	324	316
SSEQ	12	0.604	9.403	179	227	168
KonCept512	1,536	0.921	4.154	682	395	288
Virtual best method	NA	0.978	2.069	2,015	0	0

set and a test set of 8,058 and 2,015 images, respectively. Seven well-known IQA methods (BIQI, BLIINDS-II, etc.) and a newly developed deep learning method (KonCept512) were applied to the test set and gave results, summarized in Table I. More details and the references for the methods can be found in [4]. The second column of the table lists the number of features used for each method.

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 I 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 1 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 [7]. This tool automatically determines a well-performing algorithm selection approach and its hyper-parameters. 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

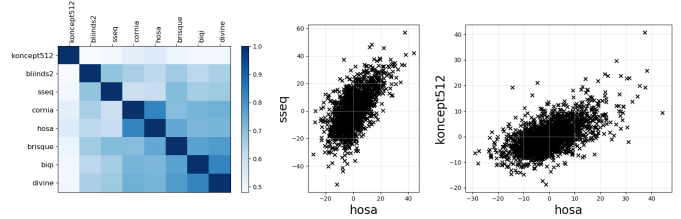


Fig. 1. Left: The correlations (SROCC) between the predictions of the 8 selected methods, clustered with Ward’s hierarchical method, are shown color coded. KonCept512’s performance across all test instances is the most different from the other seven, while others are more related, such as CORNIA and HOSA, and BIQI and DIIVINE. Right: Two scatter plots showing the (signed) errors $M(I) - MOS(I)$ for two pairs of methods. Points clustered along the vertical axis imply that the method plotted on the horizontal axis has smaller errors, and vice versa. So HOSA is more accurate than SSEQ, but less than KonCept512. Figures were generated using ASAPy [6].

TABLE II
PERFORMANCE OF SINGLE BEST METHOD (SBM), VIRTUAL BEST METHOD (VBM), AND ALGORITHM SELECTION (AS) BY AUTOFOLIO ON THE KONIQ-10K TEST SET. THE FIRST TABLE PART SHOWS THE NUMBER OF INSTANCES COVERED BY EACH METHOD.

Method	Using all methods			KonCept512 excluded		
	SBM	VBM	AS	SBM	VBM	AS
BIQI	–	187	0	–	263	51
BLIINDS-II	–	185	0	–	277	32
BRISQUE	–	176	0	–	256	140
CORNIA	–	217	0	–	329	512
DIIVINE	–	169	0	–	241	252
HOSA	–	220	0	2015	363	918
SSEQ	–	179	0	–	286	110
KonCept512	2015	682	2015	–	–	–
MAE	4.154	2.069	4.154	6.792	3.063	6.665
SROCC	0.921	0.978	0.921	0.805	0.954	0.784

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 I, 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 II 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,

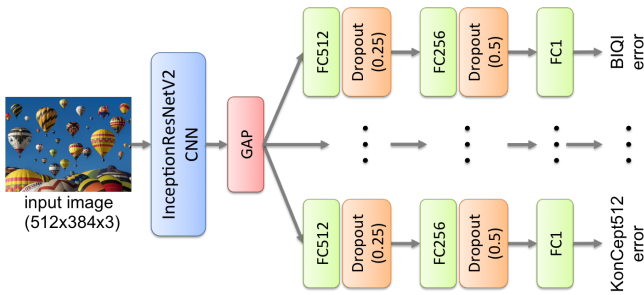


Fig. 2. The proposed siamese network architecture for error prediction.

and not for SROCC. Moreover, the large gap to the VBM’s performance (i.e., when considering just these seven) has remained.

IV. ALGORITHM SELECTION USING DEEP LEARNING

For our second attempt, a set of images with ground truth image quality values were used and split into a subset of training images and a smaller one for validation. We tried two approaches using deep learning classifiers.

Approach 1: Train a CNN-based deep learning system to classify images according to which IQA method achieves the best image quality prediction. Thus, we consider eight classes, one for each method. After training, this network provides a solution to the algorithm selection problem. We have used InceptionResNetV2 [8] for the image classification problem. Given the training set and a validation set (1,000 images, split from the training set), we fine-tuned InceptionResNetV2 with the pre-trained weight on ImageNet dataset [9], where the stochastic gradient descent optimizer was applied with a small learning rate $\alpha = 0.0001$. We trained 10 epochs with a batch size of 64 and reported the model that gave the best prediction results on validation set. This gave an image classification accuracy of 29.3% and an SROCC of 0.871 on the KonIQ-10k test set after algorithm selection.

Approach 2: For IQA methods M and images I , having ground truth quality values $MOS(I)$, we consider the error function $f_M(I) = |M(I) - MOS(I)|$. We tried to learn these functions by a Siamese neural network with a joint CNN base for all considered methods M . Then the algorithm selection for a given input image would first run this network and then output the IQA $M(I)$ of the method M , for which the network predicted the smallest error $f_M(I)$. The proposed architecture is shown in Fig. 2. We feed an image into the CNN base of InceptionResNetV2 and use Global Average Pooling (GAP) for each feature map. The resulting feature vector passes through 8 separate modules, each one predicting the error for one of the eight methods. Each module consists of five layers. These are of type fully-connected (FC) with 512 units, dropout with rate 0.25, FC with 256 units, dropout with rate 0.5, and output with one neuron. We replaced the cross entropy loss by mean absolute error loss and applied the same training process as in Approach 1. The model that gave the lowest loss on the validation set was accepted. For an

input image I it produces estimates $\hat{f}_M(I)$ of the error $f_M(I)$ for all methods M , leading to the algorithm selection result $M^*(I) = \min_{M \in \mathcal{P}} \hat{f}_M(I)$. The MAE $f_{M^*}(I)$ on the test set was 6.447, which gave an SROCC of 0.908.

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. Frech ette, 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.