OPTIMAL FEATURE SELECTION FOR BLIND SUPER-RESOLUTION IMAGE QUALITY EVALUATION

Juan Berón[⋆] *Hernán Darío Benítez Restrepo*[⋆] *Alan C. Bovik*[†]

[⋆] Pontificia Universidad Javeriana-Seccional Cali
[†]The University of Texas at Austin

ABSTRACT

The visual quality of images resulting from Super Resolution (SR) techniques is predicted with blind image quality assessment (BIQA) models trained on a database(s) of human rated distorted images and associated human subjective opinion scores. Such opinion-aware (OA) methods need a large amount of training samples with associated human subjective scores, which are scarce in the field of SR. By contrast, opinion distortion unaware (ODU) methods do not need human subjective scores for training. This paper presents an opinion-unaware BIQA measure of super resolved images based on optimally extracted perceptual features. This set of features was selected using a floating forward search whose objective function is the correlation with human judgment. The proposed BIQA method does not need any distorted images nor subjective quality scores for training, yet the experiments demonstrate its superior quality-prediction performance relative to state-of-the-art opinion-unaware BIQA methods, and that it is competitive to state-of-the-art opinion-aware BIQA methods.

Index Terms— Image quality assessment, super resolution, no reference image quality assessment

1. INTRODUCTION

Single image super resolution (SISR) algorithms aim to construct a higher resolution image based on a single image of lower resolution. Then, these types of algorithms must predict missing information between pixels. In the last two decades there have been numerous SISR methods proposed. The relative performances of these have typically been evaluated using image quality assessment (IQA) models like the peak signal to noise ratio (PSNR) and the structural similarty index (SSIM) [1].

Previous studies [2] have shown that PSNR and SSIM do not correlate very well with human perception of super-resolved image quality. Other types of metrics such as information fidelity criterion (IFC) [3] correlate better with human perception when evaluating super resolved images. PSNR, SSIM and IFC are full-reference image quality assessment (FR-IQA) algorithms, and require an original reference image against which to determine the quality of an image, which in practice is sometimes impossible to obtain. By contrast, blind image quality assessment (BIQA) algorithms do not require an original image to assess the quality. BIQA of super resolved images has received some renewed attention after the introduction of a database created by Ma *et al* [4], composed of 1620 super resolved images and corresponding subjective quality scores provided by human subjects. Although there are other annotated data sets for SR, such as [5] and [6], these are not publicly available. The Ma *et al* database has helped drive the development of several BIQA algorithms for super-resolved images, including models based on two-stage random-forest regression models [4] and convolutional neural networks [7] and [8]. These approaches are opinion-aware BIQA methods, which require a large number of super-resolved images with human subjective scores on which to learn a regression model, which can cause them to have rather weak generalization capability. Furthermore, when applying a model trained on one database to another database, the quality prediction performance can be impaired. Given these disadvantages of OA BIQA methods on super-resolved images, it is of great interest to create 'opinion-unaware' models which are not trained on samples of distortions, nor on human subjective scores. To the best of our knowledge, ODU BIQA models have not been developed to evaluate super-resolved images. This paper aims to develop an opinion-unaware BIQA method, based on an optimal feature selection process that can compete with OA BIQA methods.

The rest of this article is organized as follows. In Section 2 the perceptual quality aware features are introduced. Section 3 presents a BIQA measure that does not require training, and studies which features are best suited for the assessment of the quality of super resolution images. Section 4 analyses the performance results of the proposed metric and also studies the performance of the selected features in an OA framework as proposed in [4]. Finally, Section 5 concludes the paper.

2. PERCEPTUAL QUALITY AWARE NSS FEATURES

A promising area in BIQA is analyzing the natural scene statistics (NSS) of the images to be evaluated. NSS describe

statistical regularities in images captured by an optical camera, as opposed to machine generated images. Previous works on infrared images [9] and fused visible and long wave infrared (LWIR) images [10] have used a set of perceptual 138 features based on the processing models: mean subtracted contrast normalized (MSCN) [11], paired products [11], paired log-derivatives [12] and steerable pyramid responses [13]. In [14] and [15], the authors developed ODU BIQA metrics referred to as NIQE and IL-NIQE, based on NSS features that correlate highly with human quality perception. These techniques extract a set of local features from an image, then fit the feature vectors to a multivariate Gaussian (MVG) model. The quality of a test image is then predicted by a statistical distance between its MVG model (local or global) and the MVG model learned from on corpus of pristine naturalistic images. Our work is inspired by NIQE and IL-NIQE, nonetheless it performs better because we rely on: (i) a set of 297 enriched features obtained by combining the features deployed in [10], [4], and [16] as shown in Table 1, and (ii) an optimal feature selection process based on floating forward search, whose objective function is correlation with human judgment. This allows us to identify a set of features that are particularly sensitive to the visual distortions that occur on super resolved images.

3. IDENTIFYING AN OPTIMUM FEATURE SET

A pristine image model was created on 289 images: 170 images were taken from the BSD200 [17] image set, 29 images were selected from LIVE IQA [18] and 90 images were extracted from the database used to build the pristine model for IL-NIQE [15]. An MVG model was fitted to the 297 features extracted from the pristine images, yielding a mean vector μ and covariance matrix Σ using the standard maximum likelihood estimation method [19]. The distance to the pristine model is defined as:

$$D(x) = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{c_i^2}}, \qquad (1)$$

where $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$, $(c_1, c_2, \ldots, c_n)$ is the diagonal of Σ, n is the number of features, and $x = (x_1, x_2, \ldots, x_n)$ are the features extracted from the image whose perceptual quality is to be predicted. Super resolution reconstruction processes are complex, and a super-resolved image may contain multiple interacting distortions. Hence, as in the case of authentically distorted images [16], some perceptual features are not reliable for super resolved images.

Therefore, we deployed sequential forward floating selection [20] to select a best performing subset of features from among the 297 features. To test this subset of features, it was necessary to separate a set of images before implementing the selection procedure, so the testing set would not be correlated with the training set. This set, that we will refer as the excluded set henceforth, was made of all the super resolved versions of 8 original high resolution (HR) images selected randomly from Ma *et al* dataset. This excluded set remained fixed throughout the feature selection process. This group consisted of 432 images (8 images $\times$ 9 SR algorithms $\times$ 6 spatial resolutions), while 1188 images were used to find the best performing subset. The sequential sub-optimal forward floating selection procedure looks for the subset that maximizes the linear correlation coefficient (LCC) between the distance (1) and the perceptual scores. The procedure compares the obtained value against other possible subsets of the same size, and selects the one with the highest LCC. The subset found with the procedure is not certain to be the best, yet finds a good subset without exhaustively exploring every subset of a specific size. Through this procedure, possibly the best subset at each size is obtained. Since it is possible that the optimum set selected on each size set could end up being content dependent, the procedure was performed 100 times on 648 images selected randomly from the 1188. For each run, a maximum value of correlation was obtained, however sets whose correlation differed from the optimum one by no more than 0.02 were retained. On each run, a group of sets were selected in such a way that none were a subset of another, and each group of sets differed from the maximum correlation by 0.02. Finally, the number of times a feature appeared was counted and the final set was selected as the group of features that appeared 99% of the time. This final set included 112 features distributed as shown in Table 1.

# feats initial set	# feats final set	Description
18	3	MSCN
48	5	Paired Products
42	0	Paired log-derivatives
36	18	Steerable Pyramid
18	0	DCT
45	2	Wavelet Coefficients
75	75	PCA analysis
9	8	Sigma field
6	1	Difference of Gaussians

Table 1. Summary of the initial (297 features) and final feature space (112 features) after feature selection process.

We used the t-SNE technique [21] to visualize the high dimensional optimal set of features (using MatLab routines 'tsne' and 'seuclidean' distance) extracted from the 1620 images of the Ma *et al* database and 289 images in the pristine set. This result allows to see the tendency that when features are closer to the pristine set, the better the perceptual scores given to the image. These are encouraging results supporting the idea of using the distance from the pristine set as a quality function.

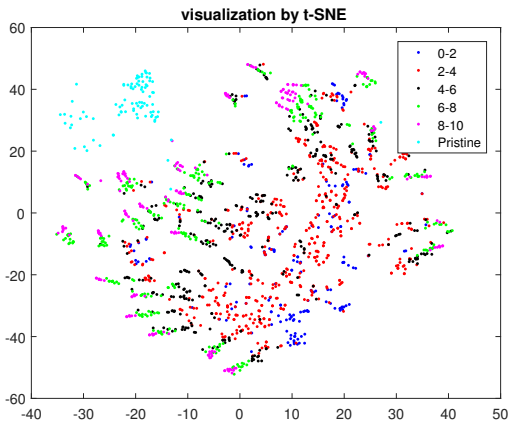

Fig. 1. t-SNE visualization using the 'seuclidean' distance. Each dot is an image, and the colors represent the range where the perceptual score of the image falls. The images of the pristine set do not have scores.

A closer look at the optimal feature set allows us to identify certain peculiarities. Specifically in regards to the features extracted from the paired products and steerable pyramid coefficient distributions. The paired product features were calculated on 3 different resolutions of each image: the original resolution, half the resolution and a quarter of the resolution. It is interesting to notice that for the optimum set, the paired product features selected were only selected at half and quarter resolution. The only features selected from the steerable pyramid set were the variances of the distributions of each subband, and they were selected at all three resolutions.

4. RESULTS AND ANALYSIS

4.1. Opinion Distortion Unaware Quality Analyzer

Our opinion and distortion unaware quality analyzer is defined as the distance between 112 dimensional representation of the pristine model, and that of the super-resolved image to be evaluated. This measures is compared against NIQE and IL-NIQE, because they are state-of-the-art ODU BIQA models. For fair comparison, the pristine image set used by NIQE and IL-NIQE was defined as the pristine set of 289 images described previously. The results of full reference IQA algorithms were also added to give a more complete view of the relative performances of the models. The linear correlation coefficient (LCC) and the Spearman rank correlation coefficient (SRCC) on the excluded set of 432 images were calculated, with the results shown in Table 2. The model defined in the 112 features dimensional space outperformed the SRCC values of IL-NIQE, NIQE, and SSIM.

Table 3 presents the results of using different combinations of features. These results indicate that when all the features are used the correlation suffers, indicating that some of the features do not correlate well with human judgments. The results obtained using the paired products at only half resolution and quarter resolution provided a better performance than using all the paired products features, which could indicate that these features performance is correlated with the resolution of the images. Furthermore, the paired products at half and quarter resolution yield improvement in performance compared with using all the features. Similarly to [14], performance vs the size of the pristine set was tested. The size of the pristine set was modified by randomly deleting a number of images. This procedure was done 1000 times, recording the SRCC each time. The results are depicted in Figure 2, indicating that even with only 20 images in the pristine set it is possible to obtain an SRCC higher than 0.8.

	LCC	SRCC
Ours	0.826	0.827
IL-NIQE [15]	0.683	0.715
NIQE[14]	0.694	0.675
SSIM [1]	0.738	0.724
IFC [3]	0.826	0.859

Table 2. Performance of different IQA algorithms on the excluded images.

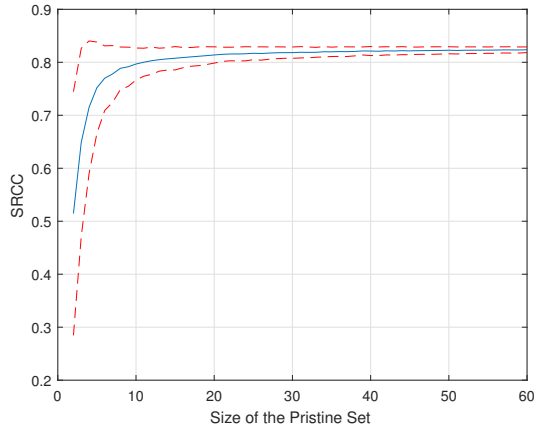

Fig. 2. Variation of SRCC vs the size of the pristine set. The blue line is the mean value over 1000 variations at each size, while the dashed red lines indicate one standard deviation confidence bands.

4.2. Opinion Distortion Aware Quality Analyzer

Once the features are selected, a common practice in image quality assessment model design is the use of a regression model. Two common regression techniques in BIQA for SR are the random forest regression (RFR) and support vector regression (SVR). RFR has shown a good performance on the

	LCC	SRCC
feats-297	0.698	0.688
Ma et al	0.703	0.673
PP-all	0.627	0.617
PP-3/4	0.766	0.790
SP-all	0.423	0.442
SP-var	0.673	0.665

Table 3. Comparison of models performances. The model feats-297 comprises the features described in Table 1. Ma et al has 138 features described in [4], all the pairwise products (PP-all), the pairwise products at half and quarter resolution (PP-3/4), all the steerable pyramid features (SP-all) and the variances of the steerable pyramid features at all resolutions (SP-var)

		Training Set Size			
		20%	40%	60%	73.3%
Ma et al	SRCC	0.860	0.886	0.893	0.896
	LCC	0.888	0.907	0.914	0.916
	RMSE	1.088	0.991	0.964	0.954
Ours	SRCC	0.896	0.913	0.919	0.922
	LCC	0.906	0.922	0.928	0.931
	RMSE	1.011	0.929	0.891	0.875

Table 4. Performance results on the excluded set as a function of training set size.

recent work of Ma *et al* [4], who designed a two stage regression model that was also applied to the optimal set of 112 features. To assess performance, the training size was varied 50 times over 20%, 40% and 60% of the data. For fair comparison, the training set did not include the excluded set of 432 images and the performance evaluation was done on the excluded set. The size of 73.3% of the training size was also added and corresponds to 1188 images which are the total images without the excluded set. The results are shown in Table 4 and correspond to the mean performance result for the 50 iterations.

We conducted another test where the training set size was randomly selected 50 times to be 20%, 40%, 60% and 80% of the 1620 images. The testing set was composed of the remaining images. The results are depicted in Table 5 and correspond to the mean performance result for the 50 iterations. As may be seen, the optimal set of features correlates better than the Ma *et al* model using every performance measure and for all training sizes. The results obtained at 80% of the training set size yielded SRCC and LCC values comparable with those obtained by a recent convolutional neural network approach reported in [7]. Additionally, we also created a trained model using the paired products features at half resolution and at quarter of the resolution (PP-3/4) under the two-stage regression model. The performance was assessed at different training sizes following the previous procedure,

		Training Set Size			
		20%	40%	60%	80%
Ma et al	SRCC	0.859	0.891	0.907	0.917
	LCC	0.884	0.910	0.921	0.931
	RMSE	1.129	1.003	0.938	0.881
Ours	SRCC	0.902	0.923	0.934	0.939
	LCC	0.913	0.932	0.942	0.948
	RMSE	0.980	0.870	0.808	0.768
PP-3/4	SRCC	0.881	0.899	0.909	0.913
	LCC	0.898	0.917	0.925	0.929
	RMSE	1.061	0.961	0.914	0.886

Table 5. Performance results by evaluating with the remaining set of images

yielding the results in Table 5. This simple model was almost as effective as the one proposed by Ma *et al* [4].

5. CONCLUSIONS

Our proposed ODU image quality analyzer delivers higher SRCC values than IL-NIQE and NIQE on super-resolved images. The OA method uses an optimal feature set and a two-stage regression model proposed by Ma *et al* [4], providing state-of-the-art prediction performance. Additionally, it was shown that the features extracted from paired products coefficients were able to produce high quality results. The evaluation phase of this study was based on the database in [4] and as shown in [15], the performance of the OA metrics could be affected by the data set used in the training phase. Towards improving study of this issue, we are developing a subjective study of super resolved images, which will be presented in future works.

ACKNOWLEDGMENTS: The authors would like to thank NVIDIA Corporation for the donation of a TITAN X GPU used in these experiments

6. REFERENCES

[1] 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, April 2004.

[2] C. Y. Yang, C. Ma, and M. H. Yang, "Single-image super-resolution: A benchmark," in *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, Eds., Cham, 2014, pp. 372–386, Springer International Publishing.

[3] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions*

on Image Processing, vol. 14, no. 12, pp. 2117–2128, Dec 2005.

[4] C. Ma, C. Y. Yang, X. Yang, and M. H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Computer Vision and Image Understanding*, vol. 158, pp. 1 – 16, 2017.

[5] C. Ledig, L. Theis, F. Huszr, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 105–114.

[6] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 PIRM challenge on perceptual image super-resolution," *Computer Vision ECCV 2018 Workshops*, p. 334355, 2019.

[7] B. Bare, K. Li, B. Yan, B. Feng, and C. Yao, "A deep learning based no-reference image quality assessment model for single-image super-resolution," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 1223–1227.

[8] Y. Fang and C. Zhang, "Convolutional neural network for blind quality evaluator of image super-resolution," in *2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Nov 2017, pp. 28–33.

[9] T. R. Goodall, A. C. Bovik, and N. G. Paulter, "Tasking on natural statistics of infrared images," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 65–79, Jan 2016.

[10] D. E. Moreno-Villamarín, H. D. Benítez-Restrepo, and A. C. Bovik, "Predicting the quality of fused long wave infrared and visible light images," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3479–3491, July 2017.

[11] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec 2012.

[12] Y. Zhang and D. Chandler, "An algorithm for no-reference image quality assessment based on log-derivative statistics of natural scenes," *Journal of Electronic Imaging*, vol. 22, 10 2013.

[13] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, Dec 2011.

[14] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, March 2013.

[15] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, Aug 2015.

[16] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, vol. 17, no. 1, pp. 32, 2017.

[17] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int'l Conf. Computer Vision*, July 2001, vol. 2, pp. 416–423.

[18] H. R. Sheikh, Z.Wang, L. Cormack, and A. C. Bovik, "Live image quality assessment database release 2," .

[19] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag, Berlin, Heidelberg, 2006.

[20] P. Pudil, J. Novovicová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119 – 1125, 1994.

[21] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Machine Learning Research*, Nov 2008.