Journal of Imaging Science and Technology® 60(6): 060410-1–060410-10, 2016.

Image Quality Assessment by Comparing CNN Features between Images

Seyed Ali Amirshahi▲
UC Berkeley/International Computer Science Institute (ICSI), Berkeley
E-mail: amirshahi62@gmail.com

Marius Pedersen▲
The Norwegian Colour and Visual Computing Laboratory, NTNU - Norwegian University of Science and Technology, Gjøvik, Norway

Stella X. Yu
UC Berkeley/International Computer Science Institute (ICSI), Berkeley

Abstract. *Finding an objective image quality metric that matches the subjective quality has always been a challenging task. We propose a new full reference image quality metric based on features extracted from Convolutional Neural Networks (CNNs). Using a pre-trained AlexNet model, we extract feature maps of the test and reference images at multiple layers, and compare their feature similarity at each layer. Such similarity scores are then pooled across layers to obtain an overall quality value. Experimental results on four state-of-the-art databases show that our metric is either on par or outperforms 10 other state-of-the-art metrics, demonstrating that CNN features at multiple levels are superior to handcrafted features used in most image quality metrics in capturing aspects that matter for discriminative perception.* © *2016 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2016.60.6.060410]

INTRODUCTION

Image quality assessment, both subjective[1] and objective,[2] is an active field of research. Objective image quality assessment methods, commonly referred to as image quality metrics, have the goal of being correlated with perceived image quality. An impressive amount of image quality metrics have been proposed in the literature,[2] yet evaluation shows that there is still room for improvement,[3] especially when it comes to performing well across databases and distortions.[2] Several researchers have pointed out challenges that are still unsolved,[4,5] such as dealing with geometric changes, multiple distortions, run-time performance and memory requirements.

Image quality metrics have several advantages over subjective assessment; they are consistent, less time consuming, and can be used for quality optimization. It is therefore important to have an image quality metric which is highly correlated with the subjective evaluation of observers.

Many different approaches for measuring image quality have been proposed, including structural similarity,[6,7] color difference,[8] spatial extensions of color difference formulas,[9–12] simulation of detail visibility,[13,14] scene statistics,[15] low- and mid-level visual properties,[16] saliency,[17] machine learning,[18–21] and more. Image quality metrics have been used to measure general image quality, but are also applied to different applications such as printing,[22–25] displays,[26,27] spectral imaging,[28] image compression,[29] and medical imaging.[30,31]

In this article, we introduce a new full reference objective image quality metric which is based on comparing different feature maps extracted from the test and reference images using Convolutional Neural Networks (CNN). CNNs have been used in many computer vision and image processing tasks. Until now most image and video quality metrics focused on comparing images using a limited number of handcrafted features[6,7,32–34] while our approach not only take low-level features into account but it also compares mid- and high-level features providing a more precise and accurate metric. Using the AlexNet model[35] we calculate the quality of the test image at different convolutional layers and link the values to provide a single score representing the overall quality of the image (Figure 1). The metric is extensively evaluated on several state-of-the-art databases and results are compared to the most accurate and famous image quality metrics introduced so far.

This article is organized as follows: Previous Image Quality Metrics reviews previous image quality metrics, Proposed Image Quality Metric describes our proposed approach, Experimental Setup Details our experimental setup, Results and Discussion reports experimental results, and Conclusion concludes.

PREVIOUS IMAGE QUALITY METRICS

Objective image quality metrics can be divided into three categories:

▲ IS&T Members.

Received July 18, 2016; accepted for publication Oct. 8, 2016; published online Dec. 13, 2016. Associate Editor: Rita Hofmann-Sievert.
1062-3701/2016/60(6)/060410/10/$25.00

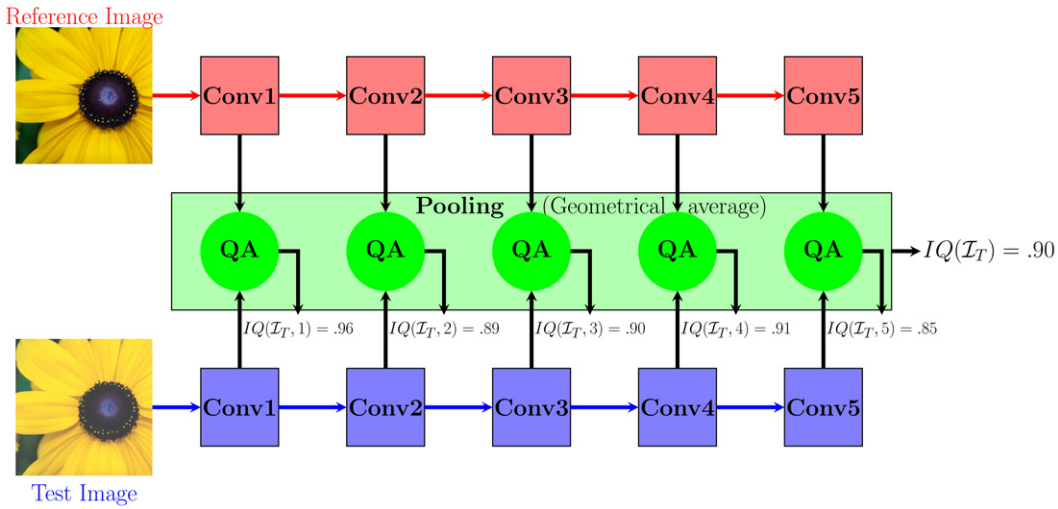


Figure 1. Our new image quality metric uses CNN features across multiple levels to compare the similarity between the test and reference images. We use the AlexNet[35] CNN model trained on the ImageNet dataset.[36] Feature maps at all five convolutional layers are extracted and compared with a histogram-based quality metric. IQ($\mathcal{I}_T$, n) denotes the image quality value at convolutional layer n (Figure 2). Their geometrical mean produces the final single value IQ($\mathcal{I}_T$) as the quality of the test image with respect to the reference image.

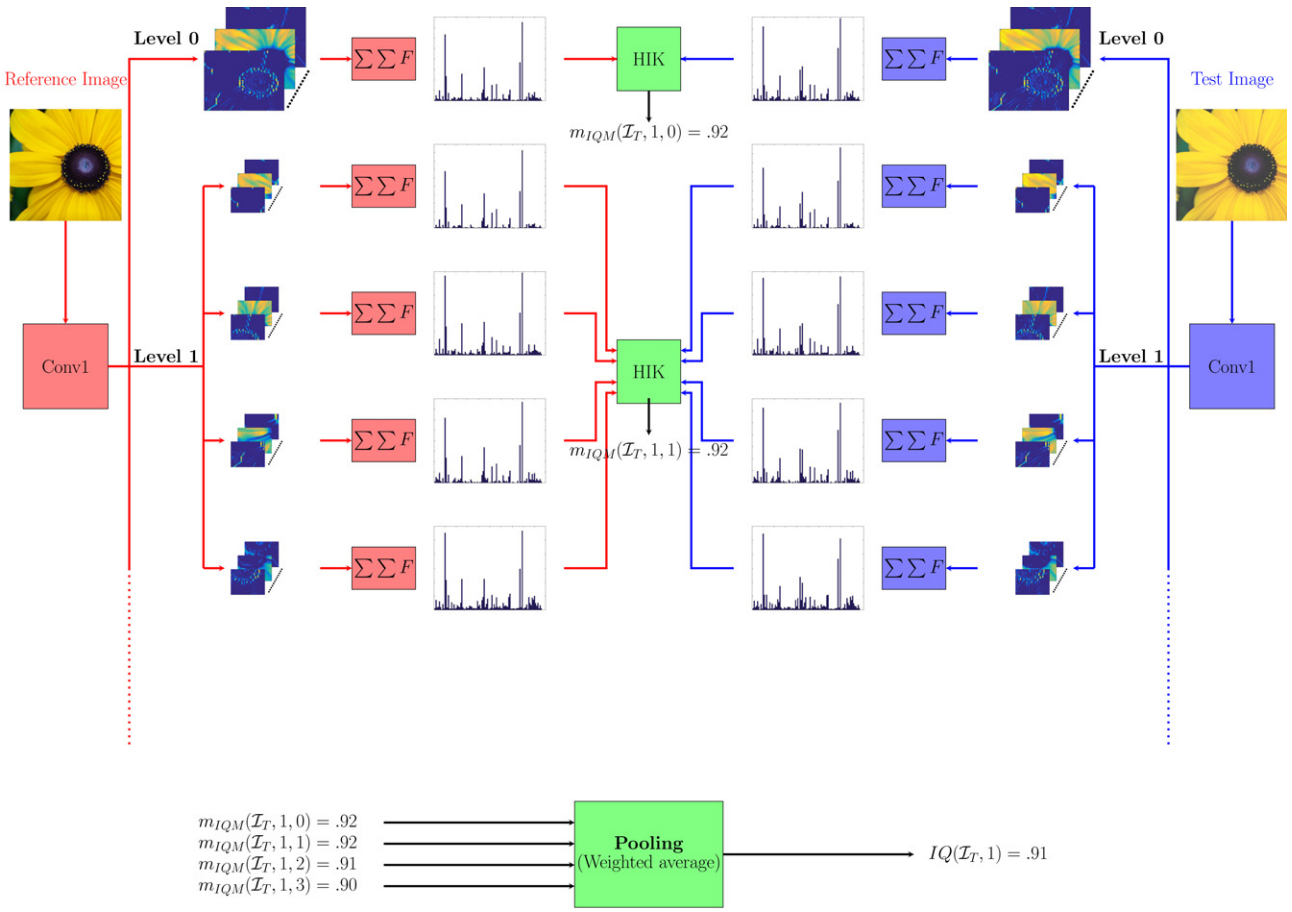


Figure 2. Pipeline used to calculate image quality at the first convolutional layer. Feature maps of the test and reference image extracted from the convolutional layers are compared to each other using the Histogram Intersection Kernel (HIK)[37] at different levels of the spatial resolution. Due to space restrictions the calculations are only shown for the ground (level zero) and first level of the spatial pyramid. The values calculated at each level are then pooled to provide a single score for each layer (shown at the bottom of the figure).

(1) **Full reference metrics**: we have access to both the reference and test images.
(2) **Reduced reference metrics**: we have access to the test image and have only partial information about the reference image.
(3) **No reference metrics**: no information about the reference image is available, but the entire test image is available.

In this section we focus on full reference metrics and review existing state-of-the-art metrics. A more complete review is available at Refs. 2, 38–40.

Full Reference Image Quality Metrics

Different full reference image quality metrics have been proposed for gray scale and color images.

Full Reference Image Quality Metrics for Gray Scale Images

The structural similarity (SSIM) index proposed by Wang et al.[7] defines the structural information as those attributes that represent the structure of the objects in the scene, independent of the average luminance and contrast. SSIM is based on a combination of luminance, contrast, and structure comparison. Comparisons are done for local windows, and the overall image quality is the mean of the local windows.

Ponomarenko et al.[41] introduced an image quality metric based on PSNR and local contrast. The metric divides the image into 8×8-pixel nonoverlapping blocks. Quality is calculated by using PSNR, and is weighted based on contrast sensitivity functions and contrast masking.

The feature similarity (FSIM) index was proposed by Zhang et al.[42]. FSIM is based on the principle that the human visual system perceives an image mainly on its salient low-level features. Two kinds of features are employed, the phase congruency and the gradient magnitude. FSIMc is the color version of FSIM.

Full reference image quality metrics for color images

Wang and Hardeberg[11] proposed a metric based on adaptive bilateral filters (ABF). The metric is based on the human visual system, where it blurs the image based on the viewing distance. The quality calculation is based on the ΔE^*_{ab} color difference formula.

Zhang and Wandell[12] proposed the S-CIELAB metric which was a spatial extension of the ΔE^*_{ab} color difference formula. The images are filtered using contrast sensitivity functions simulating the human visual system. Then the quality is calculated using the ΔE^*_{ab} color difference formula.

The spatial hue angle metrics (SHAME and SHAMEII) were introduced by Pedersen and Hardeberg.[33,34] They are based on the same framework as S-CIELAB, but incorporate a weighting based on hue. The images are filtered with contrast sensitivity functions, then the hue angle algorithm[43] is applied to account for the fact that systematic errors over the entire image are quite noticeable and unacceptable.

The iColor-Image-Difference (iCID) metric is inspired by SSIM, and was proposed by Lissner et al.[44] The metric is designed specially to improve the prediction of chromatic distortions such as those created by gamut-mapping algorithms.

CNN-based Metrics

There already exists CNN-based image quality metrics. Kang et al.[20] proposed a no reference image quality metric using the average score of CNN quality estimates for all the patches in the image. Evaluation showed high correlation with perceptual scores.

Bianco et al.[18] proposed DeepBIQ metrics, which estimates image quality by averaging the scores predicted on multiple sub-regions of the image. The score of a sub-region is calculated using a Support Vector Regression (SVR) machine over CNN features. Evaluation showed that DeepBIQ performed well compared to other no reference metrics.

Self-similarity

Recent studies in the field of computational aesthetics have proposed different approaches to measure the degree of self-similarity seen in images and especially paintings.[45–49] To evaluate the degree of self-similarity seen in an image, the mentioned methods take a pyramidal approach in which the gradient orientation seen in different regions are compared to smaller sub-regions in the image in a pyramid format. The Histograms of Oriented Gradients (HOGs)[50] is used for comparing the gradient orientations seen in the image. In this study, we extended this work to evaluate the similarity seen between two given images, the test and the reference image. The similarity between the two images was then used as a measure to evaluate the quality of the test image.

PROPOSED IMAGE QUALITY METRIC

In the proposed image quality metric, using the AlexNet[35] model which was pre-trained on the ImageNet dataset[36] implemented in the MatConvNet toolbox,[51] we evaluate the quality of images. Similar to the Pyramid of Histograms of Orientation Gradients (PHOG),[52] we compared different feature maps extracted from the test and reference image at different convolutional layers in various spatial levels (Figure 2). We expect that a test image with similar strength of features seen in different convolutional layers and spatial levels to the reference image would be similar to the reference image and show high quality values. In other words, we use the strength of the feature maps as bin entries in the pyramidal approach (similar to the PHOG method) to evaluate the similarity between two given images.

The following steps are taken in the calculation of the proposed image quality metric:

(1) As mentioned earlier, the proposed image quality metric was based on comparing the feature maps extracted at the five different convolutional layers. To compare the different histograms, the response of feature maps at different scales (levels) were compared to one another.

The first step in this approach was to calculate histogram

$$\begin{aligned}\boldsymbol{h}(\mathcal{I}_T, n, L) &= \Bigg(\sum_{i=1}^{X}\sum_{j=1}^{Y} \mathcal{F}(\mathcal{I}_T, n, L, 1)(i,j), \\ &\sum_{i=1}^{X}\sum_{j=1}^{Y} \mathcal{F}(\mathcal{I}_T, n, L, 2)(i,j), \ldots, \\ &\sum_{i=1}^{X}\sum_{j=1}^{Y} \mathcal{F}(\mathcal{I}_T, n, L, z)(i,j), \ldots, \\ &\sum_{i=1}^{X}\sum_{j=1}^{Y} \mathcal{F}(\mathcal{I}_T, n, L, M)(i,j) \Bigg), \end{aligned} \tag{1}$$

for the test image ($\mathcal{I}_T$) at level L of the spatial pyramid of the n^{th} convolutional layer. In Eq. (1), feature map z in the n^{th} convolutional layer of image $\mathcal{I}_T$ at level L is presented by $\mathcal{F}(\mathcal{I}_T, n, L, z)$ and M corresponds to the number of feature maps in convolutional layer n (in the case of the AlexNet model, 96 for the first, 256 for the second, 384 for the third, 384 for the fourth, and 256 for the fifth convolutional layer). In the equation it is assumed that the feature maps have a size of X by Y and as seen in Eq. (1) we used all the feature map size in our calculations. The sum of the response of each feature map at a given layer corresponds to the bins in histogram $\boldsymbol{h}$. The use of all feature maps in a given layer results in a better performance in our approach compared to the previous methods which tried to evaluate the similarity between two images or the self-similarity seen in an image only based on a limited number of features which were related to the gradient orientation seen in the image such as Refs. 46, 47, 49. It should be pointed out that the feature maps are extracted just after the Rectified Linear Units (ReLU) and before the max-pooling layers.

(2) To take a pyramid approach, and similar to Refs. 45, 47, each feature map was divided to four equal sub-regions (Figure 3). Histogram $\boldsymbol{h}$ (Eq. (1)) was then calculated for the new sub-regions.

(3) To maintain the pyramidal nature of our approach we continued the division of the sub-regions and calculated histogram $\boldsymbol{h}$ until the smallest side of the smallest sub-region was equal or larger than seven pixels. Keeping in mind the size of different feature maps at the five convolutional layers in the AlexNet model resulted in the third level for the first convolutional layer, the second level for the second layer, and the first level for the third, fourth, and fifth layers.

(4) Using the HIK,[37] the quality of the test image ($\mathcal{I}_T$) at level L of the spatial pyramid for the n^{th} convolutional layer was calculated by,

$$\begin{aligned} m_{IQM}(\mathcal{I}_T, n, L) &= d_{HIK}(\boldsymbol{h}(\mathcal{I}_T(n, L)), \boldsymbol{h}(\mathcal{I}_R(n, L))) \\ &= \sum_{i=1}^{n} \min(h_i(\mathcal{I}_T, n, L)), h_i(\mathcal{I}_R, n, L)). \end{aligned} \tag{2}$$

(a) Original image (b) Level 1 (c) Level 2

Figure 3. A sample photograph (a) along with its corresponding sub-regions at level one (b) and two (c) of the spatial pyramid.

From the equation it is clear that $m_{IQM}(\mathcal{I}_T, n, L)$ evaluates how similar the response of different feature maps at the given level for the specific layer of the test image ($\mathcal{I}_T$) is compared to the reference image ($\mathcal{I}_R$). Previously, different works such as Refs. 45–49, 52 used histograms that represent the strength of different features in the image to compare how similar two images are. As mentioned earlier, and pointed out using different equations, in this study we compare the strength of feature maps at different convolutional layers and spatial levels to calculate the similarity between two images.

(5) For each convolutional layer n in the test image, we introduced the quality vector

$$\begin{aligned}\boldsymbol{m}_{IQM}(\mathcal{I}_T, n) = (\, & m_{IQM}(\mathcal{I}_T, n, 1), m_{IQM}(\mathcal{I}_T, n, 2), \\ & \ldots, m_{IQM}(\mathcal{I}_T, n, z), \ldots, m_{IQM}(\mathcal{I}_T, n, L)), \end{aligned} \tag{3}$$

which is the result of the concatenation of $m_{IQM}(\mathcal{I}_T, n, l)$ values for all the levels in the spatial pyramid.

(6) Finally, the quality of the test image $\mathcal{I}_T$ at the n^{th} convolutional layer is calculated by

$$\begin{aligned}\text{IQ}(\mathcal{I}_T, n) = & \frac{1 - \sigma(\boldsymbol{m}_{IQM}(\mathcal{I}_T, n))}{\sum_{l=1}^{L} \frac{1}{l}} \\ & \times \sum_{l=1}^{L} \frac{1}{l} \cdot m_{IQM}(\mathcal{I}_T, n, l). \end{aligned} \tag{4}$$

In Eq. (4), $\sigma(\boldsymbol{m}_{IQM}(\mathcal{I}_T, n))$ corresponds to the standard deviation among the values in $\boldsymbol{m}_{IQM}(\mathcal{I}_T, n)$. The average weighting used in Eq. (4) is similar to the approach taken by Amirshahi et al.[45,47] in their measure of weighted self-similarity. The reasoning behind such weighting was to give higher importance to larger sub-regions in the image. The higher (lower) the quality of the larger sub-regions in the image is, the better (worse) the quality of the image would be. To also take into account the changes at different levels in the image we used $\sigma(\boldsymbol{m}_{IQM}(\mathcal{I}_T, n))$. This allowed us to take into account the quality changes at different levels. By using this approach, we were able to differentiate between images that have similar quality values at different levels of the spatial pyramid and images that their quality values change at different spatial levels. Through different examples, Amirshahi et al.[47] showed how using such an approach could increase the accuracy of

their self-similarity measure significantly. Finally, the $\sigma(\boldsymbol{m}_{IQM}(\mathcal{I}_T, n))$ value is subtracted from one in order to have quality scores in a manner that higher values represent better quality and low values represent low quality scores.

(7) To link the quality values calculated at different convolutional layers, we used the geometric mean

$$\mathrm{IQ}(\mathcal{I}_T) = \prod_{n=1}^{5} \mathrm{IQ}(\mathcal{I}_T, n). \tag{5}$$

This was mainly due to the fact that quality values at different layers were calculated at various spatial levels. In Eq. (5), $\mathrm{IQ}(\mathcal{I}_T)$ corresponds to the overall quality of the test image ($\mathcal{I}_T$). It should be pointed out that other than geometrical mean, we also tested other pooling approaches introduced in Ref. 53 such as the Minkowski pooling, but results did not change dramatically.

EXPERIMENTAL SETUP

There are a number of existing databases specifically created for evaluation of image quality metrics. Evaluation of the proposed image quality metric is carried out using the following databases:

- Colourlab Image Database: Image Quality (CID:IQ):[54] The database contains 23 original images modified with six distortions: JPEG 2000 compression, JPEG compression, blur, Poisson noise, ΔE gamut mapping and SGCK gamut mapping. For each distortion five levels, from low quality to high quality, were created. 17 observers rated the images in two different sessions, one session with a viewing distance of 50 cm and one session with a viewing distance of 100 cm. Ambient illumination was approximately 4 lux. The chromaticity of the white displayed on the color monitor was D65 and luminance level of the monitor was 80 cd/m^2. All settings are suited for sRGB color space.
- LIVE Image Quality Assessment Database release 2 (LIVE2):[55,56] The database contains 29 original images modified with five distortions: JPEG compressed images (233 images), JPEG 2000 compressed images (227 images), Gaussian blur (174 images), White noise (174 images), and Fast fading Rayleigh noise (174 images) resulting in a total number of 982 images. The level of distortion resulted in images at a broad range of quality, from imperceptible distortions to highly distorted images. The observers viewed the images at a distance of 2–2.5 times the screen height.
- Computational and Subjective Image Quality (CSIQ):[57] The database consists of 30 original images modified with six types of distortions: JPEG compression, JPEG 2000 compression, global contrast decrements, additive pink Gaussian noise, and Gaussian blurring. There are a total of 866 distorted images in CSIQ. Each distortion has four to five different levels of distortion. The viewing distance was approximately 70 cm. All settings are suited for sRGB color space. 35 observers participated in the subjective experiment.
- Tampere Image Database (TID2013).[58] The database contains 25 original images modified with 24 distortions: Additive Gaussian noise, Additive noise in color components is more intensive than additive noise in the luminance component, Spatially correlated noise, Masked noise, High frequency noise, Impulse noise, Quantization noise, Gaussian blur, Image denoising, JPEG compression, JPEG 2000 compression, JPEG transmission errors, JPEG 2000 transmission errors, Noneccentricity pattern noise, Local block-wise distortions of different intensity, Mean shift (intensity shift), Contrast change, Change of color saturation, Multiplicative Gaussian noise, Comfort noise, Lossy compression of noisy images, Image color quantization with dither, Chromatic aberrations, and Sparse sampling and reconstruction. Each distortion has 5 levels resulting in a total of 3000 distorted images. 971 observers participated in the experiments.

As a performance measure, we calculate the Pearson correlation between subjective scores and the scores from the image quality metrics. Pearson's correlation coefficient assumes a normal distribution in the uncertainty of the data values and that the values are ordinal. Confidence intervals as calculated using Fisher's Z-transform,[59] giving us a 95% confidence interval for the correlation values.

We calculated the linear correlation between the subjective scores and the metric scores. We also report the correlation using nonlinear regression by applying a mapping function[60]

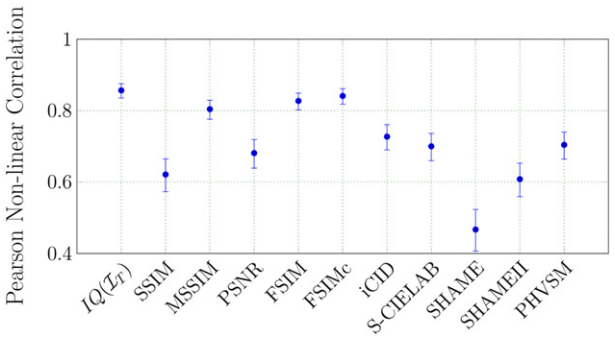
$$f(x) = \theta_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\theta_2 (x - \theta_3)}} \right) + \theta_4 + \theta_5. \tag{6}$$

Where $\theta_i, i = 1, 2, 3, 4,$ and 5 are the parameters to be fitted. Initial parameters are $max(subjectivescores)$, $min(subjectivescores)$, $median(metricscores)$, 0.1, and 0.1 respectively. To prevent repetition we only present nonlinear values to the reader.

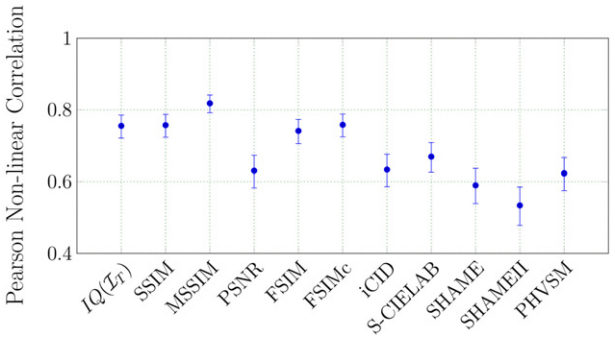
In addition to Pearson correlation, we also calculated Spearman and Kendall correlation coefficients. Results showed similar correlation rates close in value and order of performance to that of Pearson; therefore, we will only report on the Pearson coefficients.

RESULTS AND DISCUSSION

Overall the proposed metric performs very well on all datasets, always being among the best metrics. It is also stable in its performance, being a clear advantage over other metrics which perform well in one dataset while under-performing in another. In this section we will first go through the performance of the proposed image quality metrics on each dataset. We then have a look at how the proposed approach performs on images affected by different types of distortions. Then, another state-of-the-art CNN model (VGG[61]) is tested on the mentioned datasets. We also investigate how



(a) CID:IQ 100 cm



(b) CID:IQ 50 cm

Figure 4. Nonlinear Pearson correlation values for different image quality metrics calculated for the CID:IQ dataset shown with 95% confidence intervals.

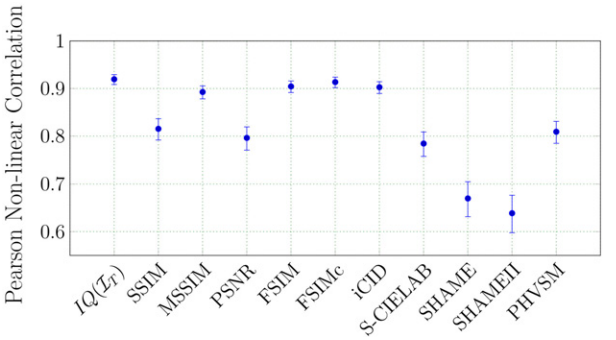


Figure 5. Nonlinear Pearson correlation values for different image quality metrics calculated for the CSIQ dataset shown with 95% confidence intervals.

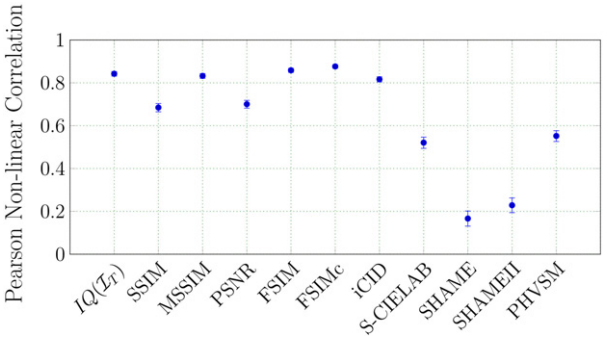


Figure 6. Nonlinear Pearson correlation values for different image quality metrics calculated for the TID2013 dataset shown with 95% confidence intervals.

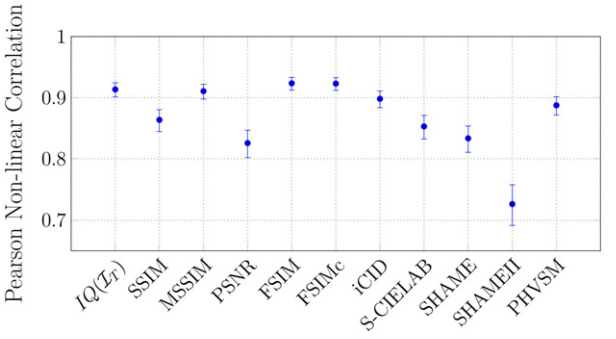


Figure 7. Nonlinear Pearson correlation values for different image quality metrics calculated for the LIVE2 dataset shown with 95% confidence intervals.

increasing the number of convolutional layers affect the performance of our proposed approach. Finally, we evaluate how quality values at different convolution layers change.

In the case of the **CID:IQ** dataset (Figure 4), compared to 10 state-of-the-art metrics, the proposed approach shows the highest Pearson nonlinear correlation in the case of the 100 cm viewing distance. For the 50 cm viewing distance, although the proposed approach is not the best, it is still among the top three performing image quality metrics. It should be pointed out that in the case of the 50 cm distance, other than the SSIM, FSIM, FSIMc, and MSSIM image quality metrics which show close correlation rates, the proposed metric performs significantly better than the six other metrics.

Among different tested image quality metrics, the proposed approach has the highest nonlinear Pearson correlation in the case of the **CSIQ** dataset (Figure 5). While the results are not significantly better compared to metrics such as FSIM, FSIMc and iCID it outperforms other state-of-the-art metrics.

Figure 6 shows the nonlinear Pearson correlation for the **TID2013** dataset. We can notice that the proposed metric performs among the best, and is significantly better than SSIM, PSNR, S-CIELAB, iCID and other image quality metrics.

Finally, in the case of the **LIVE2** dataset (Figure 7), the proposed metric has one of the highest nonlinear Pearson correlations. As shown in the figure, the results are significantly better compared to metrics such as SSIM, iCID and S-CIELAB.

Tables I, II and III provide the nonlinear Pearson correlation results for different distortions. It can be observed that the proposed image quality metric is always among the top three metrics tested in our experiments. It is interesting to observe that among all the subsets the proposed approach is the best metric in 71% of the cases which proves the high accuracy of the metric.

With regards to the quality scores at different convolutional layers ($IQ(\mathcal{I}_T, n)$, Eq. (4)), it is interesting that in general the lowest scores are observed in the case of the first layer (Figure 8 and Table IV). Keeping in mind that features in the first layer are assumed to be related to the gradient

Table I. Nonlinear Pearson correlation calculated using different image quality metrics for different distortions seen in the CSIQ dataset. In each row the highest correlation is shown by red, the second highest by blue and the third highest by green.

	IQ($\mathcal{I}_T$)	SSIM	MSSIM	PSNR	FSIM	FSIMc	iCID	S-CIELAB	SHAME	SHAME II	PHVSM
All images	**0.92**	0.82	0.89	0.80	0.90	0.91	0.90	0.78	0.67	0.64	0.81
Gaussian blurring	**0.95**	0.90	0.87	0.91	0.89	0.89	0.94	0.92	0.58	0.92	0.92
Global contrast	**0.96**	0.85	**0.96**	0.94	0.92	0.93	**0.96**	0.92	0.80	0.83	0.94
JPEG	**0.98**	0.947	**0.98**	0.89	**0.98**	**0.98**	0.97	0.97	0.92	0.91	0.97
JPEG 2000	**0.98**	0.92	**0.98**	0.95	**0.98**	**0.98**	0.96	0.95	0.87	0.66	**0.98**
Additive pink Gaussian noise	0.94	0.93	0.95	0.95	0.93	0.94	**0.96**	0.94	0.62	0.79	**0.96**

Table II. Nonlinear Pearson correlation calculated using different image quality metrics for different distortions seen in the LIVE2 dataset. In each row the highest correlation is shown by red, the second highest by blue and the third highest by green.

	IQ($\mathcal{I}_T$)	SSIM	MSSIM	PSNR	FSIM	FSIMc	iCID	S-CIELAB	SHAME	SHAME II	PHVSM
All images	0.91	0.86	0.91	0.83	**0.92**	**0.92**	0.90	0.85	0.83	0.73	0.89
Blur	**0.98**	0.87	0.96	0.78	0.97	0.97	0.93	0.83	0.87	0.91	0.92
Fast fading Rayleigh	0.91	**0.95**	**0.95**	0.89	**0.95**	**0.95**	0.94	0.82	0.80	0.74	0.89
JPEG 2000	**0.96**	0.94	**0.96**	0.90	**0.96**	**0.96**	0.95	0.90	0.87	0.81	0.95
JPEG	0.94	0.94	0.94	0.85	**0.95**	**0.95**	0.94	0.92	0.91	0.88	0.94
White noise	**0.99**	0.98	0.98	**0.99**	0.97	0.98	0.97	0.98	0.96	0.92	**0.99**

Table III. Nonlinear Pearson correlation calculated using different image quality metrics for a few of the distortions seen in the TID2013 dataset. In each row the highest correlation is shown by red, the second highest by blue and the third highest by green.

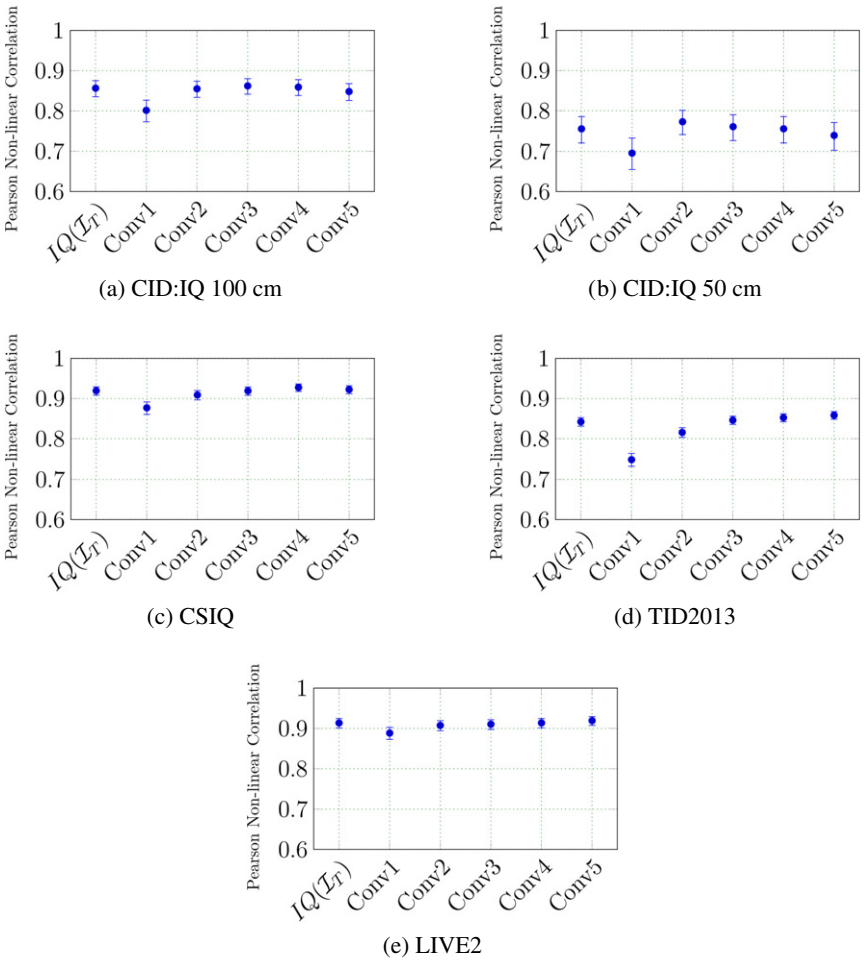
	IQ($\mathcal{I}_T$)	SSIM	MSSIM	PSNR	FSIM	FSIMc	iCID	S-CIELAB	SHAME	SHAME II	PHVSM
All images	0.84	0.68	0.83	0.70	0.86	**0.88**	0.82	0.52	0.17	0.23	0.55
Additive Gaussian noise	0.85	0.68	0.81	0.71	0.85	**0.87**	0.81	0.53	0.24	0.31	0.69
JPEG compression	0.88	0.71	0.85	0.69	0.88	**0.91**	0.79	0.51	0.67	0.26	0.69
JPEG 2000 transmission errors	**0.88**	0.66	0.77	0.67	0.78	0.81	0.81	0.60	0.28	0.55	0.64
Mean shift (intensity shift)	**0.93**	0.69	0.86	0.72	0.91	0.92	0.80	0.66	0.26	0.32	0.71
Lossy compression of noisy images	**0.88**	0.72	0.87	0.68	0.85	**0.88**	0.84	0.52	0.15	0.32	0.65
Image color quantization with dither	**0.90**	0.72	0.85	0.72	0.88	**0.90**	0.84	0.54	0.20	0.31	0.70

orientations seen in the image, we can conclude that such features do not have a high impact on the overall subjective image quality scores. Correlation scores increase in the case of the mid-convolutional layers (layers two, three and four) while the highest scores are mostly seen in the case of the last (fifth) layer. From the mentioned finding it can be assumed that while patterns and textures seen in the image (layers two, three and four) play a significant role on the overall image quality but the content of the image itself is the most influential issue when evaluating the quality of an image. This itself shows how CNNs are able to help us improve the performance of our approach. The improved accuracy of our approach compared to previous metrics which are based on a limited number of handcrafted features could be related to the fact that our results are based on CNN features which are learned from the dataset and are responsive to color, pattern, texture and objects seen in the image.

Table IV. Nonlinear Pearson correlation values at different convolutional calculated for different datasets.

	IQ($\mathcal{I}_T$)	Conv1	Conv2	Conv3	Conv4	Conv5
CID:IQ 100 cm	0.87	0.80	0.86	0.86	0.86	0.85
CID:IQ 50 cm	0.76	0.70	0.77	0.76	0.76	0.74
CSIQ	0.92	0.88	0.91	0.92	0.93	0.92
TID2013	0.84	0.75	0.82	0.85	0.85	0.86
LIVE2	0.91	0.89	0.91	0.91	0.91	0.92

In our studies, apart from the AlexNet CNN model we also calculate the proposed image quality metric using the VGG model[61] both in the case of VGG 16 and VGG 19 (Table V). It should be pointed out that compared to the AlexNet model, VGG is a deeper network with 13 convolutional layers for VGG 16 and 16 convolutional layers



(a) CID:IQ 100 cm (b) CID:IQ 50 cm

(c) CSIQ (d) TID2013

(e) LIVE2

Figure 8. Nonlinear Pearson correlation values for different convolutional layers calculated for different datasets shown with 95% confidence intervals.

for VGG 19 models. The VGG model used in our experiment was also pre-trained on the ImageNet dataset. From the results it can be observed that the proposed method based on the VGG CNN models both in the case of VGG 16 and VGG 19 do not show a dramatic change in its performance compared to the AlexNet model. This finding is a good indication on how using CNNs could significantly improve the performance of the image quality metrics while the results are stable across different models. It is also interesting to observe how increasing the convolutional layers from 13 to 16 does not really increase the performance of the proposed metric.

To further investigate how increasing the number of convolutional layers could affect the performance of our metric, we calculated the proposed metric using different number of convolutional layers (Table VI). From the results it is observed that the accuracy of the proposed metric increases as the number of convolutional layers increase in the CNN model. Compared to the AlexNet model, and keeping in mind the computational costs of the approach it can be assumed that using the AlexNet model would be a better choice for calculating the proposed metric.

Finally, it should be pointed out that the computational time of the proposed method is very low. This is due to the fact that we are working with a pre-trained network and we are simply calculating different feature maps extracted at different convolutional layers.

Table V. Nonlinear Pearson correlation values calculated at different convolutional layers for different datasets using three different CNN models.

	AlexNet	VGG 16	VGG 19
CID:IQ 100 cm	0.87	0.86	0.86
CID:IQ 50 cm	0.76	0.76	0.77
CSIQ	0.92	0.94	0.94
TID2013	0.84	0.85	0.85
LIVE2	0.91	0.96	0.96

CONCLUSION

In this study we introduced a new method to evaluate the quality of a given image using a full reference approach. The proposed image quality metric is based on extracting different feature maps at the convolutional layers of an AlexNet[35] CNN model which is pre-trained on the ImageNet dataset.[36] The feature maps of the reference and test image are then compared to each other at different spatial levels

Table VI. Nonlinear Pearson correlation values calculated at different convolutional layers for different datasets using the two different VGG models. The layer values shown in the table represent the number of the first convolutional layers used in calculating $IQ(\mathcal{I}_T)$.

	VGG 16			VGG 19		
	4 Layers	8 Layers	12 Layers	5 Layers	10 Layers	15 Layers
CID:IQ 100 cm	0.80	0.85	0.86	0.82	0.85	0.86
CID:IQ 50 cm	0.71	0.76	0.77	0.73	0.77	0.77
CSIQ	0.89	0.93	0.94	0.92	0.93	0.94
TID2013	0.73	0.78	0.84	0.78	0.81	0.83
LIVE2	0.95	0.95	0.96	0.95	0.95	0.96

to reach a quality value for each of the convolutional layers. Based on the calculated nonlinear Pearson correlation values, the proposed approach outperforms the state-of-the-art image quality metrics in most datasets and distortion types. In the case that the proposed approach is not the best image quality metric, it still stands among the top three metrics proving the good precision it has.

ACKNOWLEDGMENT

This work was supported by a fellowship within the FITweltweit program of the German Academic Exchange Service (DAAD).

REFERENCES

[1] P. Zhao and M. Pedersen, "Extending subjective experiments for image quality assessment with baseline adjustments," Proc. SPIE **9396**, 93960R (2015).
[2] M. Pedersen and J. Y. Hardeberg, "Full-reference image quality metrics: Classification and evaluation," Found. Trends® Comput. Graph. Vis. **7**, 1–80 (2012).
[3] M. Pedersen, "Evaluation of 60 full-reference image quality metrics on the CID: IQ.," *IEEE Int'l. Conf. on Image Processing (ICIP), 2015* (IEEE, Piscataway, NJ, 2015), pp. 1588–1592.
[4] D. M. Chandler, "Seven challenges in image quality assessment: past, present, and future research," *ISRN Signal Processing* (2013), pp. 1–53.
[5] Z. Wang, "Objective image quality assessment: Facing the real-world challenges," Electron. Imaging **2016**, 1–6 (2016).
[6] F. Torkamani-Azar and S. A. Amirshahi, "A new approach for image quality assessment using svd," *9th Int'l. Symposium on Signal Processing and Its Applications, 2007. ISSPA 2007* (IEEE, Piscataway, NJ, 2007), pp. 1–4.
[7] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. Image Process. **13**, 600–612 (2004).
[8] CIE. Cie 199:2011 methods for evaluating colour differences in images (2011) ISBN: 978 3 902842 38 1.
[9] G. M. Johnson and M. D. Fairchild, "A top down description of s-cielab and ciede2000," Color Res. Appl. **28**, 425–435 (2003).
[10] G. Simone, C. Oleari, and I. Farup, "Performance of the euclidean color-difference formula in log-compressed osa-ucs space applied to modified-image-difference metrics," *11th Congress of the Int'l. Colour Association (AIC)* (Sydney, 2009), p. 81.
[11] Z. Wang and J. Y. Hardeberg, "Development of an adaptive bilateral filter for evaluating color image difference," J. Electron. Imaging **21**, 023021 (2012).
[12] X. Zhang and B. A. Wandell, "A spatial extension of cielab for digital color-image reproduction," J. Soc. Inf. Disp. **5**, 61–63 (1997).
[13] M. Pedersen, X. Liu, and I. Farup, "Improved simulation of image detail visibility using the non-subsampled contourlet transform," *Proc IS&T CIC21: Twenty-first Color and Imaging Conf.* (IS&T, Springfield, VA, 2013), pp. 191–196.
[14] M. Pedersen and I. Farup, "Improving the robustness to image scale of the total variation of difference metric," *Third Int'l. Conf. on Signal Processing and Integrated Networks (SPIN)* (IEEE, Piscataway, NJ, Feb 2016).
[15] H. R. Sheikh, A. C. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," IEEE Trans. Image Process. **14**, 2117–2128 (2005).
[16] D. M. Chandler and S. S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," IEEE Trans. Image Process. **16**, 2284–2298 (2007).
[17] Z. Cai, Q. Zhang, and L. Wen, *No-Reference Image Quality Metric Based on Visual Quality Saliency* (Springer, Berlin, Heidelberg, 2012), pp. 455–462.
[18] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment" (2016), arXiv preprint arXiv:1602.05531 .
[19] A. Bouzerdoum, A. Havstad, and A. Beghdadi, "Image quality assessment using a neural network approach," *Proc. Fourth IEEE Int'l. Symposium on Signal Processing and Information Technology* (IEEE, Piscataway, NJ, Dec 2004), pp. 330–333.
[20] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2014), pp. 1733–1740.
[21] J. Lina, K. Egiazarian, and C.-C. J. Kuo, "Perceptual image quality assessment using block-based multi-metric fusion (bmmf)," *IEEE Int'l. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2012* (IEEE, Piscataway, NJ, 2012), pp. 1145–1148.
[22] T. Eerola, J.-K. Kämäräinen, L. Lensu, and H. Kälviäinen, "Framework for applying full reference digital image quality measures to printed images," *Scandinavian Conf. on Image Analysis* (Springer, Berlin, Heidelberg, 2009), pp. 99–108.
[23] P. Marius, "*Image Quality Metrics for the Evaluation of Printing Workflows*," Ph.D. thesis (University of Oslo, 2011).
[24] M. Pedersen and S. A. Amirshahi, "Framework for the evaluation of color prints using image quality metrics," *Proc. IS&T CTIV2010: 5th European Conf. on Colour in Graphics, Imaging, and Vision* (IS&T, Springfield, VA, 2010), Vol. 2010, pp. 75–82.
[25] R. Slavuj and M. Pedersen, "Multichannel DBS halftoning for improved texture quality," Proc. SPIE **9395**, 93950I (2015).
[26] P. Zhao, Y. Cheng, and M. Pedersen, "Objective assessment of perceived sharpness of projection displays with a calibrated camera," *Colour and Visual Computing Symposium (CVCS), 2015* (IEEE, Piscataway, NJ, 2015), pp. 1–6.
[27] P. Zhao, M. Pedersen, J. Y. Hardeberg, and J.-B. Thomas, "Measuring the relative image contrast of projection displays," J. Imaging Sci. Technol. **59**, 30404 (2015).
[28] S. Le Moan and P. Urban, "Image-difference prediction: From color to spectral," IEEE Trans. Image Process. **23**, 2058–2068 (2014).
[29] S. A. Amirshahi and F. Torkamani-Azar, "Human optic sensitivity computation based on singular value decomposition," Opt. Appl. **42**, 137–146 (2012).
[30] I. Kowalik-Urbaniak, D. Brunet, J. Wang, D. Koff, N. Smolarski-Koff, E. R. Vrscay, B. Wallace, and Z. Wang, "The quest for'diagnostically lossless' medical image compression: a comparative study of objective quality metrics for compressed medical images," *SPIE Medical Imaging* (International Society for Optics and Photonics, 2014), p. 903717.
[31] B. Kumar, S. B. Kumar, and C. Kumar, "Development of improved ssim quality index for compressed medical images," *IEEE Second Int. Conf. on Image Information Processing (ICIIP), 2013* (IEEE, Piscataway, NJ, 2013), pp. 251–255.
[32] S. A. Amirshahi and M. C. Larabi, "Spatial-temporal video quality metric based on an estimation of QoE," *Third Int'l. Workshop on Quality of Multimedia Experience (QoMEX), 2011* (IEEE, Piscataway, NJ, 2011), pp. 84–89.
[33] M. Pedersen and J. Y. Hardeberg, "A new spatial hue angle metric for perceptual image difference," *Int'l. Workshop on Computational Color Imaging* (Springer, 2009), pp. 81–90.

[34] M. Pedersen and J. Y. Hardeberg, "A new spatial filtering based image difference metric based on hue angle weighting," J. Imaging Sci. Technol. **56**, 50501–1 (2012).

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems* (2012), pp. 1097–1105.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *IEEE Conf. on Computer Vision and Pattern Recognition, 2009. CVPR 2009* (IEEE, Piscataway, NJ, 2009), pp. 248–255.

[37] A. Barla, E. Franceschi, F. Odone, and A. Verri, "Image kernels," *Pattern Recognition with Support Vector Machines* (Springer, Berlin, Heidelberg, 2002), pp. 83–96.

[38] A. J. Ahumada, *Computational Image Quality Metrics: A Review* (1993), Vol. 24, pp. 305–308.

[39] U. Engelke and H.-J. Zepernick, "Perceptual-based quality metrics for image and video services: a survey," *3rd EuroNGI Conf. on Next Generation Internet Networks* (IEEE, Piscataway, NJ, 2007), pp. 190–197.

[40] K.-H. Thung and P. Raveendran, "A survey of image quality measures," *Int'l. Conf. for Technical Postgraduates (TECHPOS), 2009* (IEEE, Piscataway, NJ, 2009), pp. 1–4.

[41] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of dct basis functions," *Proc. Third Int'l. Workshop on Video Processing and Quality Metrics* (2007), Vol. 4.

[42] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: a feature similarity index for image quality assessment," IEEE Trans. Image Process. **20**, 2378–2386 (2011).

[43] G. Hong and M. R. Luo, "New algorithm for calculating perceived colour difference of images," Imaging Sci. J. (2013).

[44] J. Preiss, F. Fernandes, and P. Urban, "Color-image quality assessment: from prediction to optimization," IEEE Trans. Image Process. **23**, 1366–1378 (2014).

[45] S. A. Amirshahi, *Aesthetic Quality Assessment of Paintings* (Verlag Dr. Hut, 2015).

[46] S. A. Amirshahi, M. Koch, J. Denzler, and C. Redies, "PHOG analysis of self-similarity in aesthetic images," Proc. SPIE **8291**, 822911J (2012).

[47] S. A. Amirshahi, C. Redies, and J. Denzler, "How self-similar are artworks at different levels of spatial resolution?," *Symposium on Computational Aesthetics* (ACM, 2013), pp. 93–100.

[48] J. Braun, S. A. Amirshahi, J. Denzler, and C. Redies, "Statistical image properties of print advertisements, visual artworks and images of architecture," Frontiers in Psychology **4** (2013).

[49] C. Redies, S. A. Amirshahi, M. Koch, and J. Denzler, "PHOG-derived aesthetic measures applied to color photographs of artworks, natural scenes and objects," *Computer Vision–ECCV 2012. Workshops and Demonstrations* (Springer, 2012), pp. 522–531.

[50] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'05)* (IEEE, Piscataway, NJ, 2005), Vol. 1, pp. 886–893.

[51] A. Vedaldi and K. Lenc, "Matconvnet: convolutional neural networks for matlab," *Proc. 23rd ACM Int'l. Conf. on Multimedia* (ACM, 2015), pp. 689–692.

[52] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," *Proc. 6th ACM Int'l. Conf. on Image and Video Retrieval* (ACM, 2007), pp. 401–408.

[53] M. Gong and M. Pedersen, "Spatial pooling for measuring color printing quality attributes," J. Vis. Commun. Image Represent. **23**, 685–696 (2012).

[54] X. Liu, M. Pedersen, and J. Y. Hardeberg, "CID:IQ–a new image quality database," *Int'l. Conf. on Image and Signal Processing* (Springer International Publishing, 2014), pp. 193–202.

[55] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," IEEE Trans. Image Process. **15**, 3440–3451 (2006).

[56] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "Live image quality assessment database release" **2** (2005) http://live.ece.utexas.edu/research/quality.

[57] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," J. Electron. Imaging **19** (2010) 011006–011006.

[58] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Image database tid2013: Peculiarities, results and perspectives," Signal Process., Image Commun. **30**, 57–77 (2015).

[59] Video Quality Experts Group. "Final report from the video quality experts group: Validation of reduced-reference and no-reference objective models for standard definition television, phase I." Technical Report, International Telecommunication Union (2009).

[60] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," IEEE Trans. Image Process. **15**, 430–444 (2006).

[61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition" (2014), arXiv preprint arXiv:1409.1556 .