A new objective metric to predict image quality using deep neural networks

Pinar Akyazi and Touradj Ebrahimi

Multimedia Signal Processing Group (MMSPG)
Ecole Polytechnique Fédérale de Lausanne
CH 1015, Lausanne, Switzerland

ABSTRACT

Quality assessment of images is of key importance for media applications. In this paper we present a new objective metric to predict the quality of images using deep neural networks. The network makes use of both the color information as well as frequency information extracted from reference and distorted images. Our method comprises of extracting a number of equal sized random patches from the reference image and the corresponding patches from the distorted image, then feeding the patches themselves as well as their 3-scale wavelet transform coefficients as input to our neural network. The architecture of our network consists of four branches, with the first three branches generating frequency features and the fourth branch extracting color features. Feature extraction is carried out using 12 to 15 convolutional layers and one pooling layer, while two fully connected layers are used for regression. The overall image quality is computed as a weighted sum of patch scores, where local weights are also learned by the network using two additional fully connected layers. We train our network using the TID2013 image database and test our model on TID2013, CSIQ and LIVE image databases. Our results have high correlation with subjective test scores, are generalizable for certain types of distortions and are competitive with respect to the state-of-the-art methods.

Keywords: Objective image quality assessment, full reference image quality assessment, deep convolutional neural networks, discrete wavelet transform.

1. INTRODUCTION

The vast development of digital imaging and video in today's world has created an increasing need for broadcasters and service providers to present viewers content of superior visual quality. The main goal is to achieve high quality while conforming with storage and transmission constraints. Preferred compression and transmission schemes introduce artifacts in images such as blur, blocking, ringing, contrast changes and more. There may be other imperfections in images caused during capturing or rendering data, such as poor lighting conditions of the scene or in the viewing environment. Most digital applications are targeting human viewers, therefore the assessment of perceived quality by human audience is of key importance from a multimedia signal processing view. In end-to-end systems that are delivering content to the viewers, the degree of annoyance that could be introduced at the output has to be anticipated carefully. For this purpose there are quality assessment methods that provide means of analyzing the content objectively and subjectively.

Subjective quality assessment methodologies employ human subjecsts and evaluate the quality of content with respect to viewers opinion. Procedures for subjective image and video quality evaluation involve psychophysical experiments, in which a number of viewers are given a set of stimuli to be consumed in either pre-defined laboratory settings or typical environments with less controllable conditions. The subjects' ratings are recorded and processed to be presented as an indication of the quality of the viusal content. Subjective quality assessment methodologies report to what extent the content is perceptually accurate and appealing to human audience, however there are a few drawbacks that make these impractical. The psychophysical experiments are very costly, as they are time consuming in terms of design, preparation and execution. In order to ensure high generalization ability of subjective assessments and reduce content dependency of the results, the tests have to be conducted in a very large scale which makes it even more difficult to gather abundant number of scores.

Further author information: (Send correspondence to authors) E-mail: firstname.lastname@epfl.ch

Objective assessment methods employ mathematical models to evaluate the degradation and the overall quality of visual content with respect a set of given input parameters. Unlike subjective quality assessment methodologies, the results of these methods are not dependent on opinion, therefore present the advantage of not changing between individuals. Moreover, the complexity of these models are deterministic. On the other hand, assuring the reliability of these objective quality assessment metrics in terms of their correlation with the general opinion of viewers is still a challenging research and application problem.

The general interest of image quality assessment (IQA) methods is to come up with an objective image quality metric (IQM) that is able to determine the quality of an input with high accuracy, that is, with high correlation to the would-be perceived quality. Image quality assessment methods are generally classified into three categories depending on whether information from a reference is available fully or partially, or not at all. Full-reference (FR) image quality assessment methods have access to the refence image completely, while reduced-reference (RR) image quality assessment methods use certain features extracted from the reference but not the image itself. In the case of no-reference (NR) image quality assessment methods, the reference image is not available at all and the quality evaluation is not based on the reference attributes. Since NR image quality assessment methods cannot make use of auxiliary information, in practice they are more challenging compared to RR and FR image quality assessment problems. On the other hand, all three methods have different use cases depending on the availability of the reference in any given application.

Perceptually accurate image quality assessment methods benefit mostly from natural image statistics[1] or models based on human visual system (HVS),[2] where building such models in a robust fashion is a difficult task given the complexity of the HVS itself.[3] With the recent developments in machine learning, more accurate models on image representation and classification can now be achieved.[4–6] Deep neural networks, especially deep convolutional networks have been shown to learn image representations and object classes with high generalization abilities. Such models are data-driven and rely on feature extraction and regression only, and can be used to predict the perceptual quality of the input as well.[7,8] Moreover, these models can be combined with visual saliency models to boost the accuracy, by using local weights to highlight more attractive regions in images.

In this paper, we propose a deep convolutional neural network based image quality assessment method that is able to objectively predict the quality of distorted images using the reference image (FR image quality assessment) and subjective ratings. Our network uses the spatial information as well as the frequency content of the reference and distorted images. The frequency content is analyzed by applying a three-scale wavelet decomposition on the grayscale reference-distorted image patch pairs. The convolutional neural network branches are composed of 3×3 filters[4] and employ a slightly modified residual network structure.[6,9] In order to extract features from both reference and distorted images, a Siamese network structure is adopted[7,10,11] and the features of the two images are concatenated as well as the difference of these features.[7] In order to introduce as many inputs as possible to the model, at each epoch random patches are selected from reference and distorted image pairs. The final image quality is computed as a weighted linear combination of the patch qualities, where the weights are also determined using two fully connected regression layers that estimate the influence of local patches to the global quality. The Siamese architecture can also be used for RR image quality assessment, whereas a single branch of this architecture can simply be used for NR quality assessment. We trained and tested the performance of our network using images from the TID2013 database,[12] and also conducted cross database evaluation on LIVE database[13] and CSIQ image quality database.[14] We also compared our results with those of the state-of-the-art methods in terms of Pearson's linear correlation coefficient (PLCC) and Spearman's rank correlation coefficient (SROCC) and achieved competitive results.

This paper is structured as follows. Related work on image quality assessment with an emphasis on full reference frameworks is introduced in Section 2. In Section 3 we describe our proposed framework in detail and present the experiments and results in Section 4. Section 5 concludes the paper with an overview on the results and discussion as well as future directions.

2. RELATED WORK

The simplest measures to assess the quality of the distoted image in a FR framework are mean square error (MSE) and its derivatives, signal to noise ratio (SNR) and peak signal to noise ratio (PSNR). While these

metrics are widely used, they do not involve any perceptual information and do not correlate well with subjective ratings.[15] The need to include HVS responsiveness to the quality assessment yielded to other metrics, such as structural similarity index (SSIM).[16] The structural information (cross-correlation) was defined as the attributes representing the objects in a scene, and evaluated independently from the average luminance (mean) and contrast (variance) of the image. SSIM index is applied locally rather than globally due to the variances of luminance and contrast, and was further extended to multi-scale structural similarity index (MS-SSIM).[17] Feature similarity index (FSIM)[18] based on SSIM adds the comparison of low-level feature sets between the reference and the distorted images. Image fidelity criterion (IFC)[19] and visual information fidelity (VIF)[20] are other objective metrics that use natural scene statistics modeling with an image degradation model and an HVS model.

Another efficient objective metric inspired by HVS is difference of Gaussian (DOG)-SSIM[21] that computes a nonlinear combination of features extracted from several difference of Gaussian frequency bands, which mimics the contrast sensitivity function of the HVS. DOG-SSIM has feature extraction and regression steps using random forest implementation. DeepSim[8] employs VGGnet architecture[4] and computes local similarities between the features at each layer and explores various pooling methods to estimate a global quality score. Although the aforementioned methods involve feature learning and regression, to the authors' best knowledge, the only end-to-end trained method for FR image quality assessment is Deep Image Quality Measure (DIQaM-FR). This work again uses the VGGnet architecture[4] with 10 convolutional and 5 pooling layers for feature extraction combined with two fully connected layers for regression in a Siamese network. The reference and distorted images are separated into random patches to allow for artificial data augmentation. The features from reference and distorted image patches are fused before regression to increase the accuracy of the model. Another extension of DIQaM-FR is the weighted version of the model, referred to as the WaDIQaM-FR, where two fully connected layers running in parallel to the quality regression layers are added. These layers estimate the weights of local patches to the overall quality score, thereby including a saliency weighted distortion pooling.

Although the performance of DIQaM-FR and WaDIQaM-FR are superior to the state-of-the-art methods when trained and tested on the full TID2013 database, the generalization ability of the model is still limited in cross-database evaluations involving CSIQ and LIVE databases. This suggests that extracting only spatial features and followed by regression could benefit from auxiliary information such as frequency characteristics of the images in mutliple scales. Additionally, it has been shown that deep residual network architectures show better generalization abilities compared to plain VGGnet architectures, and ease the optimization by providing faster convergence at earlier stage even for very deep networks.[6]

3. PROPOSED FRAMEWORK

The proposed framework first extracts features from both the reference and distorted images. These features are then concatenated into a single feature vector that is passed onto the fully connected layers for regression and an objective quality score is assigned at the output. The architectures we have built for feature extraction, feature concatenation and regression have been inspired by the implementations in.[7] We also used a Siamese network to extract features from both the reference and distorted images, where we changed the design of the convolutional layers and the preferred building blocks.

First, we have decided to use the color channels of images as well as the wavelet decomposition of the grayscale images up to three scales as the input. 2-D wavelet decomposition is known to be effective in image processing tasks such as denoising, interpolation, sharpening and compression by providing information about both the spatial and frequency content of the image in different scales. High frequency components in images, such as edges and textures, can be distinguished from other components such as noise. The distortions in natural image databases contain artifacts such as blur, noise, illumination effects, blocking and more. These artifacts may affect distinct frequency areas of the images in a different fashion. It is therefore important to analyze the differences between both high frequency components and low frequency approximations of the referefence and distorted images as a result of distortion. Hence, we have computed the discrete wavelet transform of the reference and distorted images up to three scales using Daubechies wavelets and used the wavelet coefficients for feature extraction.

For building the convolutional layers, we were mainly motivated by the idea behind VGGnets.[4] The convolutional layers in VGGnets are the composed of kernel sizes as small as 3×3 and two basic design rules are

followed: (i) the layers have the same number of filters given an output size, and (ii) if the output size is halved then the number of filters is doubled in order to preserve the time complexity per layer. We follow a similar architecture as explained in,[6] given that the residual connections result in a model that is easier to optimize and exhibit lower training error when the depth increases.

To make maximum use of our limited training database, we divided the input images into N_p number of patches that are selected randomly. The dimensions of each patch are determined as 128×128, thereby resulting in wavelet decompositions of size 64×64, 32×32 and 16×16. Images are normalized prior to network processing. Our proposed VGGnet inspired residual convolutional layers are comprised of 8 to 10 weight layers with 3 to 4 shortcut connections for wavelet coefficient inputs and color patch inputs, respectively. The features are extracted using a series of 3x3 conv 32, 3x3 conv 32, 3x3 conv 64, 3x3 conv 64, 3x3 conv 128, 3x3 conv 128, 3x3 conv 256, 3x3 conv 256, with an addition of 3x3 conv 512, 3x3 conv 512 for the color input. The shortcut connections for residual architecture are established by 1x1 convolutional filters of size 64, 128 and 256, with an additional filter of size 512 for the color input. The downsampling is performed by using convolutional layers of stride 2 instead of pooling. At the end of each branch we used a 1x1 convolutional layer with 16 filters to reduce the output size. Further dimensional reduction is performed for the branches with input size greater than 16×16 by using max pooling. All max pooling layers have 2×2 sized kernels. The output of each branch is then concatenated to form the final feature vector, as shown in figure 2. The convolutional layers of same output size are activated through a leaky rectified linear unit (Leaky ReLU) where $\text{LeakyReLU}(x) = \max(0, x) + 0.01 \times \min(0, x)$. The choice of this activaton function is to allow a small nonzero gradient when the unit is not active, thereby preventing all outputs from reducing to zero. Instead of random initial weights, we have used the robust He initialization method that considers the rectifier nonlinearities[22] for all convolutional and linear filters.

The complete architecture of our network is depicted in Figure 1. Following the feature extraction of both reference and distorted image patches, the distorted image features f_D are concatenated with the reference image features f_R. Moreover, we also add the difference vector $f_D - f_R$ as the accuracy of the model is reported to increase by using this configuration in.[7] The feature vectors are then passed through two fully connected layers for regression, FC 256 and FC 1. Between these layers we again use the Leaky ReLU activation prior to dropout regularization with a ratio of 0.5 in order to prevent overfitting.[23] The feature vector is separately fed into two fully connected layers for computing local patch weights. The architecture of this block is the same with the output regression layer, FC 256 and FC 1. Between these layers, ReLU activation prior to dropout with a ratio of 0.5 is used. Furthermore, a final ReLU activation is applied before weight computation, in order to ensure the weights are greater than or equal to zero. Afterwards a small constant $\epsilon = 1\text{e}{-6}$ is added to the weights to prevent zero weights.

For an input patch i, the computed weight is a_i such that $a_i = \max(0, a_i^*) + \epsilon$ where a_i^* is the output prior to ReLU activation. The quality of patch i is computed in the parallel regression branch as y_i. The overall image quality is then computed as a linear combination of patch qualities and patch weights:

$$\hat{q} = \frac{\sum_i^{N_p} a_i y_i}{\sum_i^{N_p} a_i} \tag{1}$$

The loss function we have used for training our model is the mean squared error between the computed image quality and the ground truth, i.e., the MOS rating of the image. The proposed network is trained iteratively by back propagation[24,25] over a number of epochs until the error is stabilized. An epoch is defined as the period during which training takes place until the whole data has been processed by the network once. For batchwise optimization, the training data is divided into batches during each epoch. In each batch we have used two images, from which $N_p = 32$ patches are extracted. Data augmentation is carried out by flipping each image from left to right and choosing additional $N_p = 32$ patches from each of the flipped images. Therefore we have a total of 128 image patches per batch. The backpropagated loss is the average of overall losses between the computed scores and MOS values of the four images in each batch. As was done in,[7] patches are randomly sampled every epoch to introduce as many different inputs as possible to the network during training. The ADAM method[26] is used for batch optimization with the recommended parameters of $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and a decaying learning rate starting from $\text{lr} = 10^{-4}$ with a decay percentage of 10% every 5 epochs. The loss is computed on

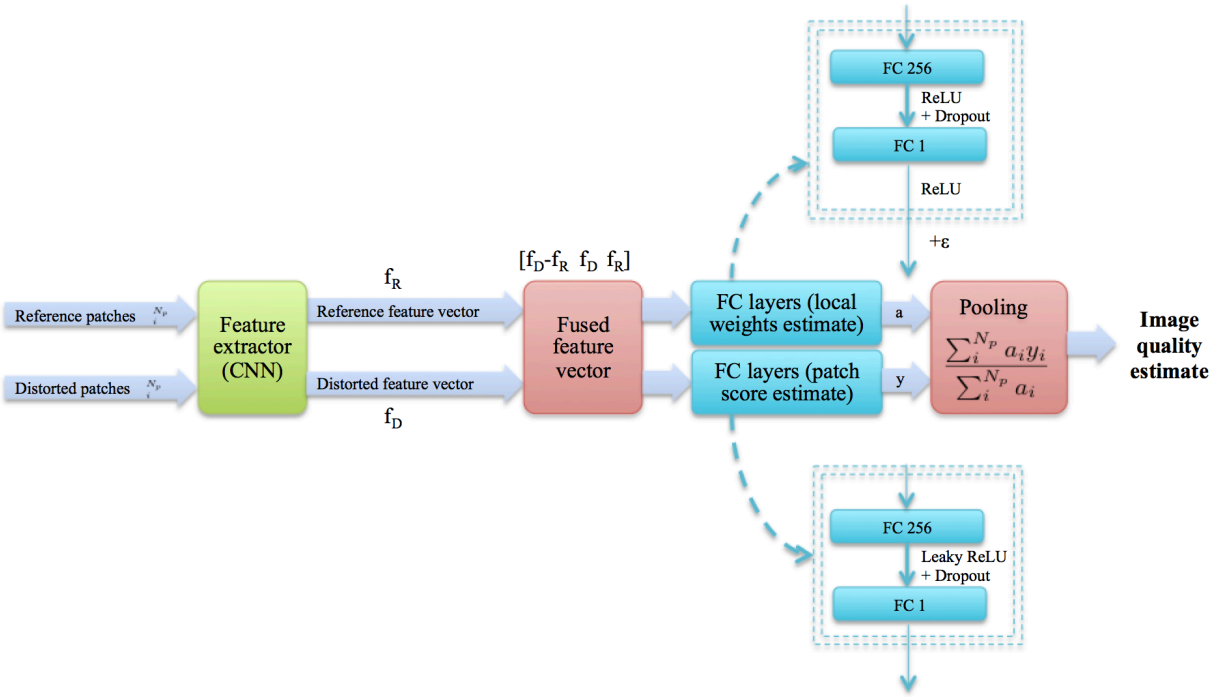


Figure 1: The proposed framework for training and testing our model. Features are extracted from both reference and distorted image patches, using color information and wavelet decomposition. The reference and distorted feature vectors are concatenated, also with a third difference vector. The final feature vector is passed through parallel fully connected layers for local weight estimation and patch score estimation. Overall score of each image is computed as a linear combination of the weighted patch scores.

a separate validation set at the end of each epoch, where the validation set is defined at the beginning of the algorithm instead of choosing random patches at every epoch in order to ensure stability. The final model used for accuracy tests is the model with least validation error, which corresponds to using early stopping criterion to stop training.[27]

4. EXPERIMENTS AND RESULTS

4.1 Datasets

Our deep neural network is trained and tested on the TID2013 database. For cross database evaluations the same network is used to compute the objective scores of images in LIVE and CSIQ databases.

The TID2013 database contains 25 reference images and 3000 distorted images, where for each reference image there are 24 types and 5 levels of distortion. A wide spectrum of distortion types have been included in this database, including additive Gaussian noise, Gaussian blur, high frequency noise, quantization noise, JPEG compression, JPEG2000 compression, lossy compression of noisy images and sparse sampling and reconstruction. MOS values of the database lie in the range $[0, 9]$ with 0 being the lowest quality score and 9 the highest. We have separated the TID2013 dataset into training, validation and test sets randomly, using 15, 5 and 5 images respectively.

The LIVE database contains 29 reference images and 779 distorted images, with 5 different distortion levels for various distortion types such as JPEG, JPEG2000, Gaussian blur, white noise and bit errors in JPEG2000

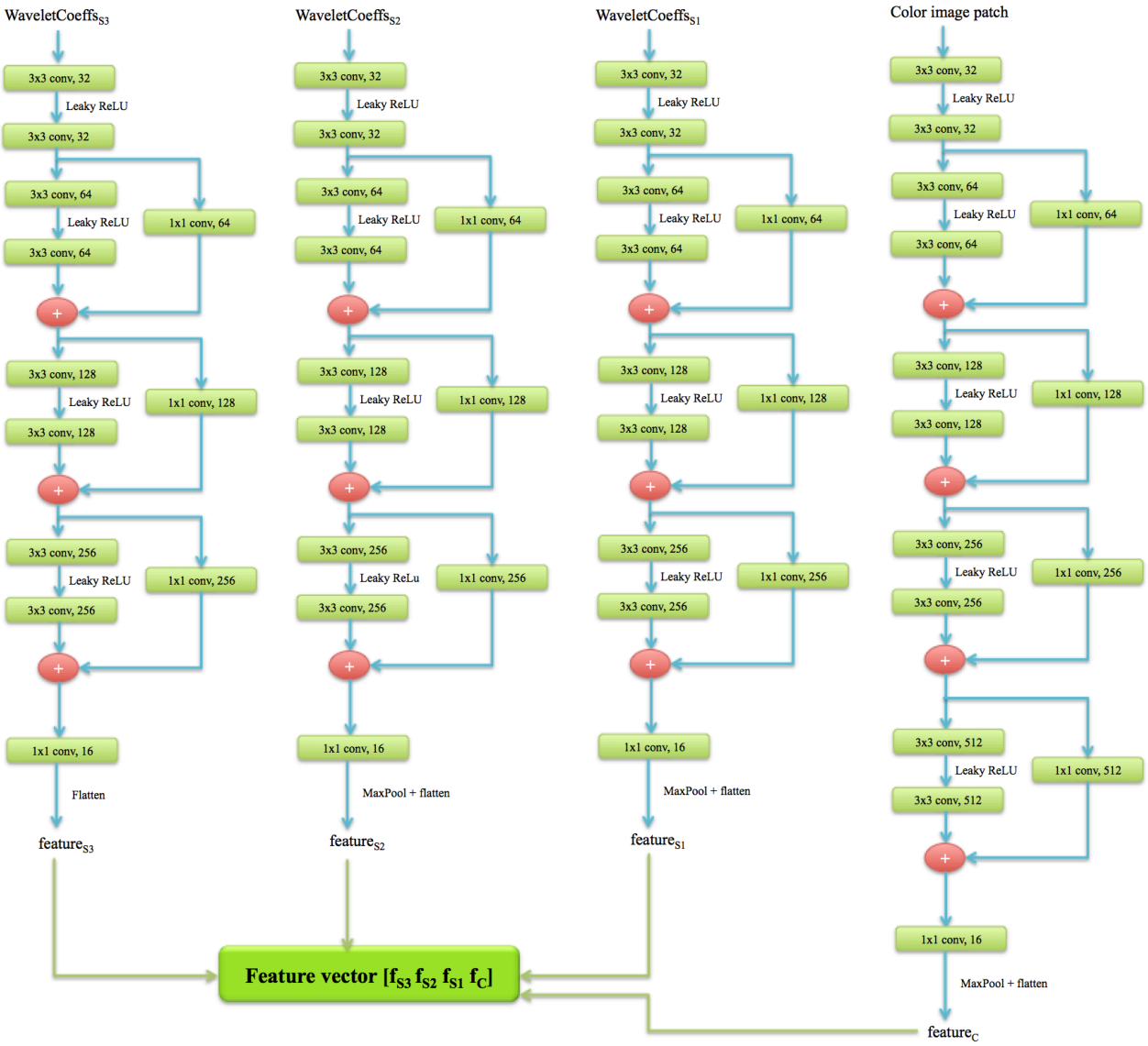


Figure 2: Feature extractor composed of convolutional layers, as previously shown in Figure 1 as the CNN block. Inputs of the first three branches from left to right are the wavelet coefficients of the 128×128 image patch, where S3 corresponds to the coarsest scale and S1 corresponds to the finest scale. The rightmost branch is the color image patch branch. Features are extracted using a VGGnet inspired architecture involving shortcut connections and 1x1 convolution at the end for dimensional reduction. Max pooling is also applied when necessary. Feature vectors of four branches are concatenated into a final feature vector of the input image patch.

bitstream. Degradation mean opinion score (DMOS) values are reported in the range [0, 100], where 0 indicates the best quality and 100 indicates the worst quality.

CSIQ image database consists of 30 reference images, each distorted at four to five different levels of distortion. The types of distortion included in the CSIQ image database are JPEG compression, JPEG2000 compression, Gaussian blur, Gaussian white noise, Gaussian pink noise and contrast change. Like in LIVE database, the ratings are reported in the form of DMOS in the range [0,1].

4.2 Results on TID2013 database

We have trained our model on the TID2013 image database using five random splits and computed the average test accuracy over the five test sets. The total number of epochs is 100 for each model, and for each test the model with the lowest validation loss has been selected. The averaged training and validation losses are shown in Figure 3. We have used the same test images to evaluate the PSNR, SSIM, FSIMc (c stands for color) and WaDIQaM-FR metrics. WaDIQaM-FR model used in our tests has been downloaded from the publicly available set of models in https://github.com/dmaniry/deepIQA, where we have chosen the model trained on the TID2013 database for consistency. Table 1 presents the performance comparison in terms of the PLCC and SROCC values with respect to the MOS values of each image, averaged over all test images.

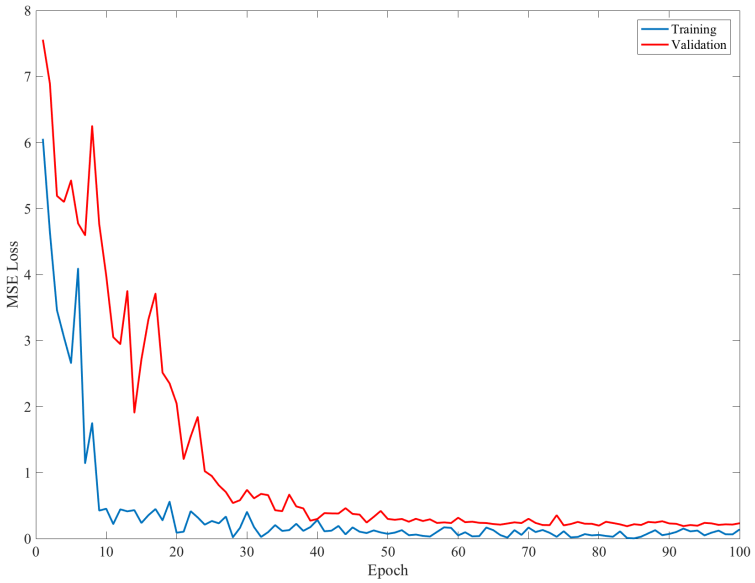


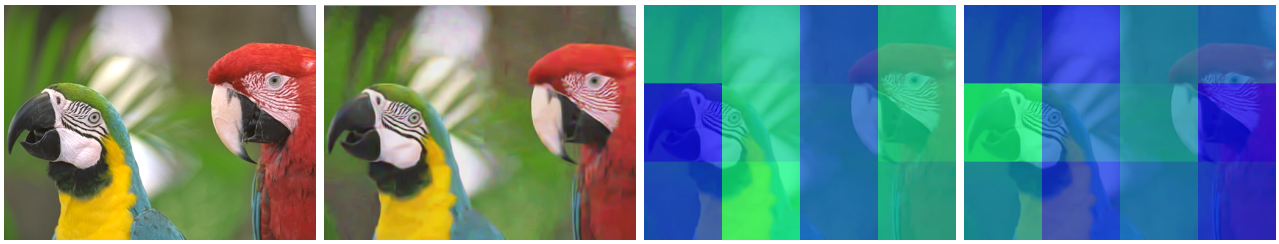
Figure 3: Training and validation losses for 100 epochs on the proposed model, averaged over five runs with random splits for the training and validation sets. Each training set has 15 reference images, whereas validation sets comprise of 5 reference images.

From Table 1 we see that the performance of the proposed method is superior to other tested methods in terms of PLCC and SROCC on the TID2013 test images. To elaborate on how an image score is determined, we have examined an example test image from the database in figure 4. The reference image is distorted due to an image denoising algorithm, where a distortion of level 4 out of 5 has been introduced. Although we have used 32 overlapping random patches per image for training, we have chosen 12 non-overlapping patches on the image in order to illustrate local patch scores and weights. In figure 4(c) we see the patch scores computed by the proposed method, where the colormap changes from blue (low) to green (high). Figure 4(d) depicts the corresponding local weights for the patches 4(c). The MOS of the distorted image in figure 4(b) is given as 3.6191 out of 9.000, which indicates a low perceptual quality. Indeed, we see that the weights assigned to patches with

Table 1: Performance comparsion of the proposed method and PSNR, SSIM, FSIMc, and WaDIQaM-FR in terms of PLCC and SROCC. The reported results have been averaged over five randomly selected tests, each consisting of 5 reference images and the corresponding 600 distorted images in the TID2013 database.

IQM	PLCC	SROCC
PSNR	0.6192	0.7113
SSIM	0.6189	0.5934
FSIMc	0.8512	0.8565
WaDIQaM-FR	0.8566	0.8488
Proposed	**0.8964**	**0.8729**

high local quality is low, and weights assigned to patches with lower quality is high, resulting in an overall low score that correlates well with the MOS.



(a) (b) (c) (d)

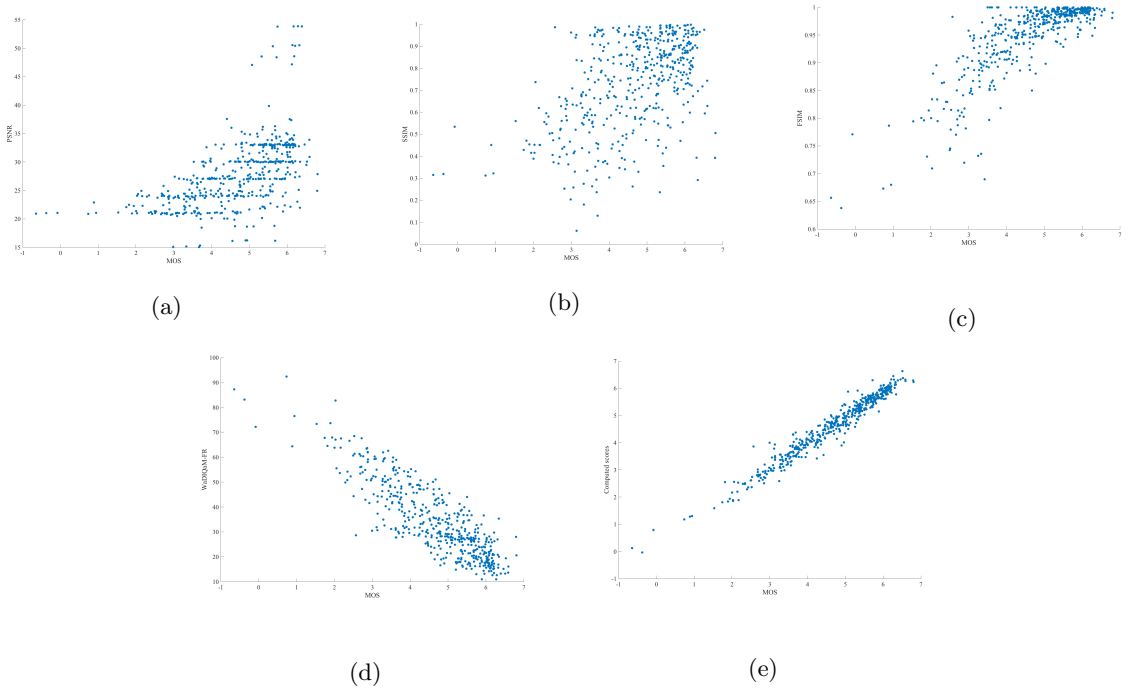
Figure 4: An example reference image from the TID2013 datased used for testing (a). The distorted image as a result of image denoising, with distortion level 4 out of 5 (b). The local scores computed by the proposed method for each non-overlapping 128×128 patch using the reference and distorted image pair, overlayed on the distorted image (c). The local weights computed by the proposed method corresponding to each non-overlapping 128×128 local score patch, overlayed on the distorted image (d). For (c) and (d), the colormap changes from blue to green, where blue indicates low and green indicates high values. Mean squared error loss between the computed score and the given MOS is 4.319×10^{-5}.

In figure 5 we have depicted the computed objective metrics versus the MOS values on the TID2013 image database test set. Comparing with the correlation coefficients presented in table 1, we can see that below 80% PLCC and SROCC, the distribution of computed scores are very random. WaDIQaM-FR and the proposed method are highly correlated with the MOS values, however we can observe that the variance of WaDIQaM-FR scores are higher than the proposed method.

4.3 Cross-database evaluation

In order to test the generalization ability of our model, we have performed objective quality assessment experiments on separate image databases containing different reference and distorted images. In tables 2 and 3, we present the performance of our method compared to that of WaDIQaM-FR on the LIVE and CSIQ image quality databases, respectively.

The full LIVE database is comprised of 779 distorted images, however 433 of these share common contents with that of the TID2013 database. In order not to bias our results, we have not included the common contents as test material for quality assessment on LIVE database. We have also scaled the DMOS values reported for the LIVE database to conform with the range we have used in training. The results depicted in table 2 have been evaluated on the remaining 346 distorted images of the LIVE database. Reported correlations concerning our method have been averaged on the five different random split models we have trained. A closer examination indicates that on the full LIVE database, the overall performance of WaDIQaM-FR is superior to our model in terms of both PLCC and SROCC. The same trend is observed for distortions caused by JPEG and JPEG2000



(a) (b) (c) (d) (e)

Figure 5: Computed objective metrics (a) PSNR, (b) SSIM, (c) FSIMc, (d) WaDIQaM-FR and (e) the proposed method on the TID2013 test images versus the MOS values.

Table 2: Performance comparsion of the proposed method and WaDIQaM-FR in terms of PLCC and SROCC on the test images selected from LIVE database.

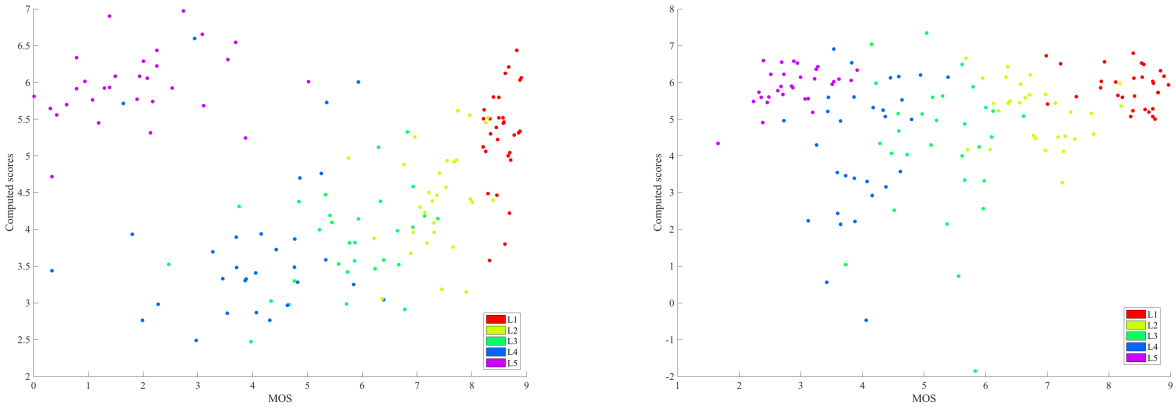
Distortion type	WaDIQaM-FR		Proposed	
	PLCC	SROCC	PLCC	SROCC
All	**0.8624**	**0.8843**	0.7182	0.8158
JPEG2000	**0.8653**	**0.8749**	0.8124	0.8404
JPEG	**0.8915**	**0.8915**	0.7691	0.7508
White Noise	0.8814	0.8732	**0.9038**	**0.9338**
Gaussian Blur	**0.9419**	**0.9374**	0.7477	0.9181
Fast Fading Rayleigh	**0.9168**	0.9191	0.8193	**0.9289**

Table 3: Performance comparsion of the proposed method and WaDIQaM-FR in terms of PLCC and SROCC on the full CSIQ image quality database.

Distortion type	WaDIQaM-FR		Proposed	
	PLCC	SROCC	PLCC	SROCC
All	**0.6261**	**0.6426**	0.5190	0.5434
Noise	**0.9309**	**0.9161**	0.8769	0.8144
JPEG	**0.9250**	0.9008	0.8769	**0.9311**
JPEG2000	**0.9186**	0.8427	0.0828	**0.8914**
Frequency noise	**0.1859**	**0.0616**	0.1219	0.0424
Gaussian Blur	**0.2970**	**0.3620**	0.0936	0.0089
Contrast	0.6990	0.9167	**0.7325**	0.7496

compression and Gaussian blur. Our results are highly correlated with the MOS values in the case of white noise, and fast fading Rayleigh, i.e. bit errors.

None of the 866 distorted images in the CSIQ image database share the same content with TID2013, hence for tests on the CSIQ image database we do not leave out any content. The CSIQ database tests are therefore more comprehensive compared to the tests on LIVE database. For the experiments, the DMOS values of the CSIQ database are again scaled to conform with our training values. Table 3 indicates that on the full database, our model has higher correlation with the reported MOS of CSIQ database only for JPEG and JPEG2000 distortions in terms of SROCC, and contrast distortions in terms of PLCC. The overall correlation of the proposed method with the underlying MOS values is much lower than WaDIQaM-FR, which is mostly brought down by the evaluations on frequency noise and even more by Gaussian blur, where reported scores exhibit a random distribution. WaDIQaM-FR correlations have also dropped down for these two types of distortions, however less drastically. In figure 6(a), we see that the ratings of our model are very high for the highest level of blur distortion in the CSIQ database. These erroneous ratings are partly due to the difference between the levels of blur distortion introduced in TID2013 and CSIQ databases, where the highest level of distortion in the CSIQ database is higher compared to the maximum blur distortion in the training data. With such elevated levels of distortion, high frequency features of the image are lost, resulting in a loss of information brought by the wavelet coefficients. In this case the blurred image can be regarded as a very smooth and therefore high quality content. A different effect is observed in the case of additive pink Gaussian noise, as depicted in figure 6(b). This type of noise has not been introduced in the training set, therefore both our results and WaDIQaM-FR have reduced correlation levels. As the level of frequency noise is increased, sharp frequency features from wavelets elevate the ratings in our model, resulting in uncorrelated scores with the underlying MOS.



(a) (b)

Figure 6: Computed objective metrics versus MOS on CSIQ database for Gaussian blur (a) and additive pink Gaussian noise (b) types of distortion. L1 indicates the lowest level of distortion while L5 indicates the highest level of distortion.

The main differences between the proposed model and WaDIQaM-FR is the inclusion of wavelet decomposition as auxiliary information for quality evaluation, and the distinct architectures of feature extraction layers. Focusing on the auxiliary information, we see that frequency characteristics of the reference and distorted images helped the model to perform fair objective quality assessment on the TID2013 database. The generalization ability of the model, however, is limited. Nevertheless we observe competitive performance of the model with the state-of-the-art in cross database results, with superior correlation values with underlying MOS values for white noise and fast fading Rayleigh distortions in LIVE database, and JPEG, JPEG2000 and contrast distortions in CSIQ database, respectively.

In order to examine the effect of the wavelet decomposition in our model, we have conducted additional experiments where the color image patch branch is removed from the architecture in figure 2. In this scenario,

only the wavelet coefficients are used for feature extraction and color information is ignored. Table 4 depicts the extent of influence brought by the wavelet coefficients to our scheme. Although the color information is undoubtedly useful for objective quality assessment, we observe that only the features extracted using the three scale wavelet decomposition are capable of capturing the quality level of the distorted image compared to the reference with high accuracy.

Table 4: Influence of wavelet coefficients in objective quality assessment for our model. The correlations are averaged over the test sets of TID2013 in five random splits.

Feature extraction	**PLCC**	**SROCC**
Wavelet coefficients	0.8572	0.8372
Wavelet coefficients + Color	0.8964	0.8729

Examining the residual architecture of feature extraction layers, we see that our preference helps the training error converge much faster compared to the WaDIQaM-FR model. In less than 40 epochs, our model is able to reach the accuracy levels presented in this paper. On the contrary, preferred WaDIQaM-FR models are expected to converge to a stable loss around 1000 epochs.[7] In terms of complexity, our model has approximately 9.5M parameters and the network experiences around 14M different training patches in 100 epochs. The complexity of WaDIQaM-FR is much lower with approximately 5.2M parameters, where the number of epochs are reported as 3000 in,[7] resulting in around 178M patches introduced to the system during training. Although our model has higher complexity, it is able to learn important features for objective quality assessment relatively faster and with less number of inputs.

In order to increase the generalization ability of our model, a thorough exploration of the hyperparameters is of key importance. This involves testing the effect of using different number of scales, patch sizes, number of patches and learning rates. Our relatively low accuracy levels on the LIVE database suggest the need to better incorporate the influence of frequency information in the feature extraction step. Our feature vectors are a concatenation of equally weighted wavelet features in different scales and the color features. Similar to learning the weight of local patches on the overall image quaility, the weights of these features can be learned to yield more descriptive feature vectors and hence more accurate results.

5. CONCLUSION

In this paper we have introduced a new full reference objective metric to predict distorted image quality using deep neural networks. Our model makes use of both the color features of the reference and distorted images, as well as the frequency characteristics extracted from a 3-scale discrete wavelet transform using Daubechies wavelets. The feature extraction module of our network is inspired by VGGnets and has shortcut connections that build residual blocks. We use a Siamese network architecture to extract features from the reference and distorted images simultaneously. Regression of patch feature vectors into global image scores is carried out using fully connected layers, computing local scores and their respective weights. Our model has been trained on the TID2013 image database, and tested on the same database as well as LIVE and CSIQ image quality databases. Results on the TID2013 database show superior prediction accuracy compared to the state-of-the-art methods. Results on cross-database evaluation indicate that for particular types of distortion our method performs better than the state of the art, however we see there is room for improvement in the interpretation of the frequency information for distortion types and levels that have not been included in the training. We observe that the frequency features help boosting the performance of our metric in the TID2013 database and yield the highest correlations with the underlying MOS ratings, but a more robust feature extraction strategy has to be followed in order to increase the generalization ability of our model on cross-database evaluations.

Future work consists of optimization of our feature extraction scheme, using regression for aggregation of features extracted from the reference and distorted image patches. Another adaptation that needs to be applied is enhancing the training set, as we have observed that the network needs to see more types and higher levels of distortion for improved generalization ability. Further improvements can be established by exploring different hyperparameters such as the learning rate decay, number of patches, patch dimensions and number of wavelet

decomposition scales. Additional methods for increasing the generalizaion ability of our model involve using a different loss function instead of the mean squared error, such as the correlation between the batch scores and the ground truth, following the same strategy for evaluation. Alternatively, the local weighting of patch scores may benefit from robust saliency models based on image quality evaluation in addition to the weighted patch aggregation model presented in our work. Following these modifications, one immediate next step is to adapt our objective image quality metric for video, making use of the temporal correlations between frames and implementing an end-to-end optimization framweork for predicting video quality.

ACKNOWLEDGMENTS

This paper reports a research performed under the framework of project Digital Eye: Deep Learning Video Quality Assessment Technology, funded by The Swiss Commission for Technology and Innovation (CTI) under the grant 27403.1 PFES-ES.

REFERENCES

1. A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE transactions on Image Processing* **20**(12), pp. 3350–3364, 2011.
2. H. R. Sheikh and A. C. Bovik, "Image information and visual quality," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, **3**, pp. iii–709, IEEE, 2004.
3. Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?," in *ICASSP*, **4**, pp. 3313–3316, 2002.
4. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* , 2014.
5. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.
6. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
7. S. Bosse, D. Maniry, K. R. Mller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing* **27**, pp. 206–219, Jan 2018.
8. F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu, "Deepsim: Deep similarity for image quality assessment," *Neurocomputing* **257**, pp. 104–114, 2017.
9. S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *CVPR*, **1**(2), p. 3, 2017.
10. J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," in *Advances in neural information processing systems*, pp. 737–744, 1994.
11. S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, **1**, pp. 539–546, IEEE, 2005.
12. N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, *et al.*, "Image database tid2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication* **30**, pp. 57–77, 2015.
13. H. Sheikh, "Live image quality assessment database release 2," *http://live. ece. utexas. edu/research/quality* , 2005.
14. E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging* **19**(1), p. 011006, 2010.
15. W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation* **22**(4), pp. 297–312, 2011.
16. 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* **13**(4), pp. 600–612, 2004.
17. Z. Wang, E. Simoncelli, A. Bovik, *et al.*, "Multi-scale structural similarity for image quality assessment," in *ASILOMAR CONFERENCE ON SIGNALS SYSTEMS AND COMPUTERS*, **2**, pp. 1398–1402, Citeseer, 2003.
18. L. Zhang, L. Zhang, X. Mou, D. Zhang, *et al.*, "Fsim: a feature similarity index for image quality assessment," *IEEE transactions on Image Processing* **20**(8), pp. 2378–2386, 2011.
19. 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* **14**(12), pp. 2117–2128, 2005.
20. H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, pp. 23–25, 2005.
21. S.-C. Pei and L.-H. Chen, "Image quality assessment using human visual dog model fused with random forest," *IEEE Transactions on Image Processing* **24**(11), pp. 3282–3292, 2015.

22. K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.
23. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research* **15**(1), pp. 1929–1958, 2014.
24. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE* **86**(11), pp. 2278–2324, 1998.
25. Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*, pp. 9–48, Springer, 2012.
26. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* , 2014.
27. L. Prechelt, "Early stopping-but when?," in *Neural Networks: Tricks of the trade*, pp. 55–69, Springer, 1998.