Image Quality Assessment Using Similar Scene as Reference

Yudong Liang, Jinjun Wang(✉), Xingyu Wan, Yihong Gong,
and Nanning Zheng

Institute of Artificial Intelligence and Robotics,
Xi'an Jiaotong University, Xi'an, China
{liangyudong,wanxingyu}@stu.xjtu.edu.cn,
{jinjun,ygong,nnzheng}@mail.xjtu.edu.cn

Abstract. Most of Image Quality Assessment (IQA) methods require the reference image to be pixel-wise aligned with the distorted image, and thus limiting the application of reference image based IQA methods. In this paper, we show that non-aligned image with similar scene could be well used for reference, using a proposed Dual-path deep Convolutional Neural Network (DCNN). Analysis indicates that the model captures the scene structural information and non-structural information "naturalness" between the pair for quality assessment. As shown in the experiments, our proposed DCNN model handles the IQA problem well. With an aligned reference image, our predictions outperform many state-of-the-art methods. And in more general case where the reference image contains the similar scene but is not aligned with the distorted one, DCNN could still achieve superior consistency with subjective evaluation than many existing methods that even use aligned reference images.

Keywords: Image Quality Assessment · Similar scene referenced image · Structural similarity · "Naturalness" · Dual-path Deep Convolution Neural Network

1 Introduction

Assessing the quality of a distorted image would benefit from the availability of a reference image. As revealed in [1], human are more skilled at comparing images than making direct judgement of the image quality. Accordingly, human can evaluate the quality of an image more accurately and consistently when provided with a high-quality reference image, and meanwhile human may give different quality scores to the same image if different reference images are presented [2]. The situation is the same for Image Quality Assessment (IQA) algorithms, where methods that make use of reference images could achieve better consistency with

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46454-1_1) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part V, LNCS 9909, pp. 3–18, 2016.
DOI: 10.1007/978-3-319-46454-1_1

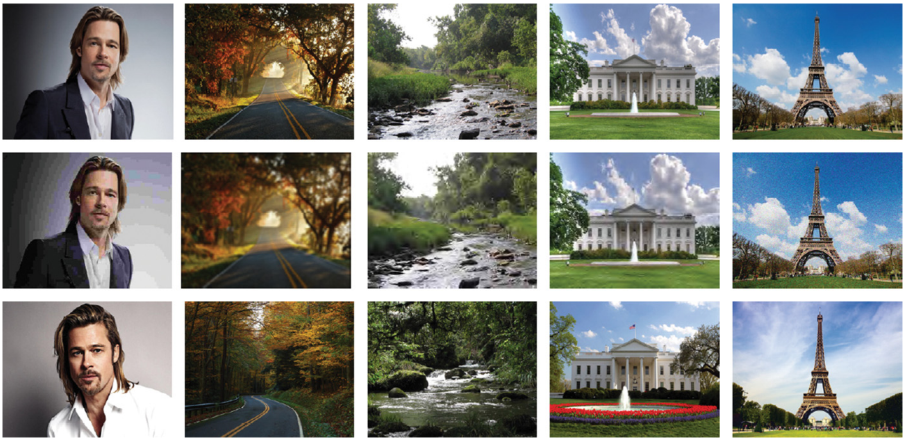


Fig. 1. Using non-aligned images with similar scene as quality reference. *Original images (top), distorted images (middle) and reference images (bottom) that are only similar to but is neither aligned with nor related by any geometrical transformation with the distorted images*

subjective assessments than those that do not consider references [3,4]. Based on whether and how reference images are used, existing IQA methods could be broadly categorized into the following three groups: full reference (FR) IQA methods [2,3,5,6], reduced referenced (RR) IQA methods [7,8], and no reference (NR) IQA methods [9,10]. The former two groups, *i.e.*,the FR-IQA and the RR-IQA methods groups, take advantage of complete or partial information of the reference image respectively, while the NR-IQA methods are often designed to extract discriminative features [9,10] or to calculate natural scene statistics to qualify the image quality [11]. As explained above, FR-IQA methods often achieve more consistent assessment as human.

However, one strong assumption with most FR-IQA methods is that, the reference image must be pixel-wise aligned with the distorted image for assessment. The requirement could be satisfied if the task is, *e.g.* to measure the quality of JPEG 2000 (JP2K) compression. Unfortunately, in more general scenario, the imaging process that generates the distorted image and the reference image may not produce aligned pair. For instance, a cell phone camera may capture a photo with hand-shake, and it is difficult to then capture an aligned high-quality image as reference. An image enhancement module on an automatic vehicle can improve a low-quality capture of the road but to assess the performance of the enhancement, it is impossible to put the vehicle at the same position to shot a pixel-aligned picture for reference. In both these common scenarios, only NR-IQA method could be used to assess the quality of the distorted images.

In this paper, we are interested in whether the image quality could be assessed using a reference image with similar scene but is not aligned, as illustrated in Fig. 1. We term the problem as NAR-IQA (Non-aligned Reference IQA).

Studies in Human Visual System (HVS) have shown that, HVS presents different sensitivity to different image signals such as spatial frequency [4], luminance [12], structural information [5], etc. Among all these features, the success of the SSIM [5] metric and its extension [13] indicates that measuring the scene structural information does benefit IQA. In addition, the visual attention property [12,14], or well known as saliency [15], tells that human usually pays attention only to a smaller but representative part of the scene, and therefore it is reasonable to assume that, if the reference image contains the same scene structure as the distorted image, it can still be used to evaluate the quality of the distorted one. Unfortunately, limited literature is available for NAR-IQA approaches. One example is the CW-SSIM [16] method that attempts using reference image with small affine transformation (scale, rotation and translation) to assess the quality of medical and binary images, which performs unsatisfactory for natural images as observed in our experiments.

This has motivated us to design a Dual-path deep Convolutional Neural Network (DCNN) for image quality assessment, using reference image of similar scene but not necessarily aligned. The two paths take the distorted image and the reference image respectively. Through weight sharing between paths, the same kind of features are extracted at the lower stage of the model. At the final stage, the proposed model concatenates features from both paths, and then a regressor is used to predict the image quality score.

Experimental results first validate that, the NAR-IQA problem is solvable where in case a pixel-wise aligned reference image is not available, a non-aligned image with similar scene can be well used as reference. In addition, our proposed model handles the IQA problem well. As explained above, the FR-IQA problem can be regarded as a special case of the NAR-IQA problem where an aligned reference image is given to the model. In this case, our predicted image quality scores are more consistent with subjective evaluations than many state-of-the-art methods. In more general case, *i.e.*,the NAR-IQA problem, our model could still achieve superior consistency than many existing methods that even use aligned reference images. Hence, while there are previous works that attempted IQA with small geometrical transformation between the reference and the distorted images [16], to the best of our knowledge, our work is the first to support IQA from reference image with similar scene but is not aligned, such that reference images become obtainable for more IQA applications.

2 Related Work

A large body of FR-IQA methods has been proposed to judge the quality of distorted images by considering reference images. Most of the approaches in this line require the reference image to be strictly aligned with the distorted images. Simple error sensitivity metrics such as the Mean Square Error (MSE) or the Peak Signal-to-Noise Ratio (PSNR) compare local pixel difference between the reference image and the distorted image, but in general the evaluation does not correlate well with human assessment. The SSIM [5] method modeled the structural information inspired by HVS for quality judgement and could achieve more consistent evaluation as human. Wang *et al.*[13] further combined multiscale information to improve the situation. Zhang *et al.* [3] carefully designed the phase congruency and gradient magnitude as complementary features of FSIM for IQA. Different from FR-IQA methods, the RR-IQA methods [7,8] focus on utilizing only parts of the reference image information to accomplish the assessment. In general, FR-IQA and RR-IQA mimic different sensitivity of HVS to different image signals [2,17], including spatial frequency, luminance [12], structural information [5] etc. to devise metric for IQA. As explained before, most FR or RR-IQA methods are extremely sensitive to small geometrical misalignment between the reference and the distorted images, which severely limits their applications.

It is worth noting that, the CW-SSIM [16] algorithm in the FR-IQA class was designed to handle very small scale, rotation and translation changes between the distorted image and the reference. On the other side, since natural images have more variations in frequency domains, in practice, the CW-SSIM method seems to perform well only for medical and binary images but poor for more general cases such as photos, surveillance footage and natural images.

Since aligned reference images are not always available, NR-IQA methods [9,10] have aroused extensive interests. Focus of most NR-IQA methods is to obtain discriminate features, and nowadays many NR-IQA algorithms are based on set of training data to learn such feature rather than proposing handcrafted ones. *e.g.*, the CORNIA [10] method learned codebooks from local image patches to encode features, and then regressor was trained to predict the quality of a distorted image. Kang *et al.*utilized CNN, a most popular model in the deep learning domain that has recently show excellent performance on visual feature learning [18], to learn features for NR-IQA [9] and achieved impressive result that approaches the state-of-the-art FR-IQA performance. This has motivated us to also apply deep learning technology for feature learning in the stated NAR-IQA problem. As demonstrated in our experiment, under the FR-IQA scenario, our predicted image quality score is more consistent with subjective evaluation than many state-of-the-art methods, and in the NAR-IQA scenario, our model still achieves superior consistency than many existing FR-IQA methods. The next section depicts our approach.

3 IQA with Non-aligned Reference

The wide availability of images from mobile phones, webcam, camcorder and Internet provides possibility to alleviate the non-existence problem of reference images for IQA. In most cases, however, reference images obtained in this way may or may not pixel aligned with the distorted image, or the two are not related by any geometrical transformation, as illustrated in Fig. 1. Hence in this section we want to propose a model that could make use of such type of reference. The model does not loose the capacity of existing FR-IQA methods but further supports the NAR-IQA problem. The proposed model is presented in the following subsection.

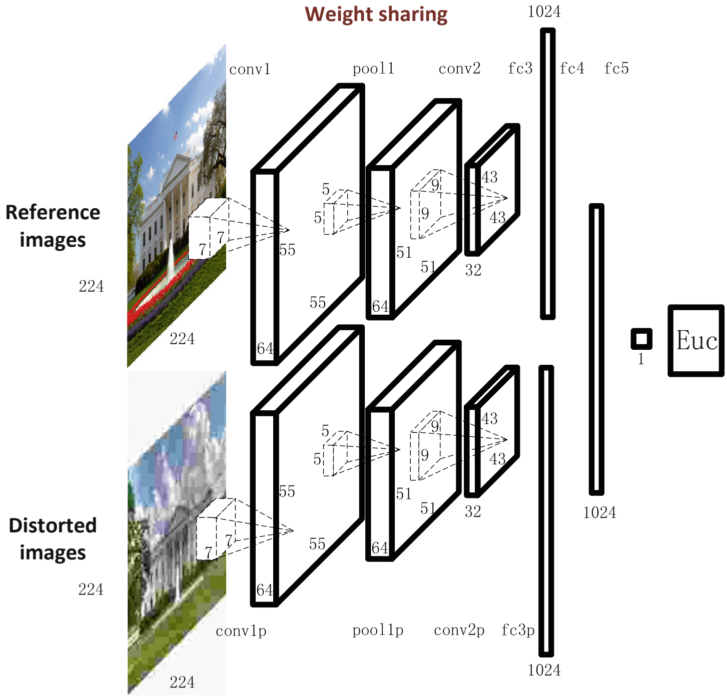


Fig. 2. The proposed Dual-path deep Convolutional Neural Network (DCNN)

3.1 Dual-Path Deep Convolutional Neural Network

We present a Dual-path deep Convolutional Neural Network (DCNN) to accept two channels of inputs and output one image quality score.

To use the proposed DCNN for NAR-IQA, we first decompose the input distorted image and the reference image into multiple standard 224×224 sub-images (Note that since input pair is not aligned, these sub-images are not necessarily aligned), then each pair of sub-images is fed to the model to obtain an quality score, and finally the overall image quality score takes the average score of all pairs of that image. The architecture of our proposed DCNN model is illustrated in Fig. 2. It consists of the convolutional layer, the nonlinear rectified linear unit, the pooling layer, the concat layer and the full connection layer, denoted as *conv#*, *relu#*, *pool#*, *concat#* and *fc#* respectively. The configuration of DCNN is listed in Table 1.

3.2 Layers

The *conv* layers are trained to extract local features. In a recursive fashion, denoting A_i^j as the feature map of path i in the j^{th} layer, W_j and B_j as the weight and bias parameters of the j^{th} layer, then the local information is extracted into deeper layers by Eq. (1), where $*$ denotes the calculation of convolution.

Table 1. Configuration of DCNN for NAR-IQA

Layer name	Padding	Filter size/stride	Output size
*input*1	0		$224 \times 224 \times 3$
*conv*1 / *conv*1p	0	$7 \times 7/4$	$64 \times 55 \times 55$
*relu*1 / *relu*1p			$64 \times 55 \times 55$
*pool*1 / *pool*1p (MAX)		$5 \times 5/1$	$64 \times 51 \times 51$
*conv*2 / *conv*2p	0	$9 \times 9/1$	$32 \times 43 \times 43$
*relu*2 / *relu*2p			$32 \times 43 \times 43$
*fc*3 / *fc*3p			1024
concat	concatenating the features from *fc*3 and *fc*3p		
*fc*4			1024
*fc*5			1
Euc			1

$$A_i^{j+1} = W_j * A_i^j + B_j, \tag{1}$$

In order to make comparison between the distorted and the reference images, we want to extract the same type of features for the two paths. Thus a weight sharing strategy is applied in the dual paths. Besides, to increases the nonlinear properties and accelerates training, the activation function is selected to be *relu* as follows: $A_i^{j+1} = max(0, A_i^j)$

Another important issue for NAR-IQA is to compensate the offset between similar scene content from the distorted image and the reference image. The *pool* layers are exploited for the purpose by integrating features from a larger local receptive field. Both [9,10] proposed to use max and min pooling over an entire feature map to align the response. Although both works achieved impressive results, these two methods abandoned the structural information which is valuable for IQA problem. Our model considers a rather large sub-image (224×224, Table 1), and we integrate the information from local to global as the network goes deeper. For computational efficiency, max-pooling is applied as Eq. (2), where R is the pooling region of corresponding position.

$$A_i^{j+1} = \underset{R}{max}\, A_i^j \tag{2}$$

The *concat* layer concatenates the features from the both paths. Then with the *fc#* layer, discriminative features are further combined and mapped to generate image quality assessment in a linear regressor.

Finally, the image quality score is predicted by minimizing the following Euclidean loss,

$$\underset{W,B}{min} \;||\big(f(I_{ref}, I_{dis}); W, B\big) - Eva||^2 \tag{3}$$

where I_{ref}, I_{dis} and Eva are the input sub reference, distorted images and human evaluations respectively, W, B are the parameters of convolutional and fc layers.

3.3 Preprocessing

Different from traditional IQA methods which need carefully designed handcrafted features, our proposed DCNN model learns discriminative features from raw data to maximally preserve information from image. Only simple local contrast normalization is needed to ensure numeric stability. The process can also be understood as a data whiten process where the intensity value of pixel $I(x, y)$ is normalized as [9],

$$\begin{aligned} I(x,y)_N &= \frac{I(x,y) - u(x,y)}{\sigma(x,y) + \epsilon} \\ u(x,y) &= \sum_{a=\frac{-P}{2}}^{a=\frac{P}{2}} \sum_{b=\frac{-Q}{2}}^{b=\frac{Q}{2}} (I(x+a, y+b)) \\ \sigma(x,y) &= \sqrt{\sum_{a=\frac{-P}{2}}^{a=\frac{P}{2}} \sum_{b=\frac{-Q}{2}}^{b=\frac{Q}{2}} (I(x+a, y+b) - u(x,y))^2} \end{aligned} \tag{4}$$

where $I(x, y)_N$ denotes values at image location (x, y) normalized by pixels in a neighboring $(P \times Q)$ window, and ϵ is a small positive constant. Although intensity shift and contrast variation sometimes were considered to be distortion, it is highly subjective to judge the image quality for distortion with these type. And we mainly deal with image distortion from degradation. Thus, Eq. (4) was applied for the input.

3.4 Training

DCNN can be trained using stochastic gradient descent with the standard backpropagation [18]. In particular, the weights of the filters of the *conv* or *fc* layer can be updated as Eq. (5)

$$\begin{aligned} \triangle_{i+1} &= m \cdot \triangle_i - \eta \frac{\partial L}{\partial W_i^j} \\ W_{i+1}^j &= W_i^j + \triangle_{i+1} - \lambda \eta W_i^j \end{aligned} \tag{5}$$

where m is the momentum factors, η is the learning rate, j is index of the layer and $\triangle_{i+1}$ is the gradient increment for training iteration i. λ is the weight decay factor. Momentum factor and weight decay factor were fixed in 0.9 and 0.0005 respectively in the following experiments.

3.5 Discussion

The architecture of the proposed model focuses on extracting features and to avoid pixel-wise aligned. This is achievable at deeper layers that integrate information from different receptive fields of earlier layers. The convolution, pooling and other nonlinear operations capture structural information from the local to the global area without explicit pixel-wise alignment, and therefore make the model geometrical robust. Also regarding the final *fc*# layers, they behave much more complicated than simple element-wise subtractions. *fc*# layers have obtained weights to not only gauge the image distortion from pairs but also ignore the feature disagreement from two paths caused by nonalignment. All the distorted samples have different image contents (such as affine variations) from the reference counterpart, thus image contents are not discriminative.

This has made the architecture and key design strategy of our proposed DCNN model very different from the well-known Siamese [19] model, which has been widely used in face verification [20]. Siamese network use contrastive loss for classification while our dual path CNN (original for IQA problem) is based on Euclidean distance to perform regression. Since our proposed model focus on regression rather than classification, in our DCNN, the mapping from the concatenated features to the image quality score is automatically learned during the training process.

4 Experiment

In this section, we report a series of experiments to validate the effectiveness of the proposed model. The Deep learning toolbox Caffe [21] was applied to built the DCNN model for IQA. Three datasets were adopted in the experiments: The LIVE dataset, the TID2008 dataset and an in-house collected dataset.

Note that since the reference images are aligned with the corresponding distorted images, in order to train and test our proposed DCNN for NAR-IQA, we synthesized the non-aligned reference image by applying affine transform to the original reference image, as nonaligned ref images preserve structures with affine transformation. The scaling factors s# and rotation θ were randomly ranged from [0.95 1.05] and [$-5°5°$] respectively. As shown in Fig. 3, a pair of training samples are collected as follows: for each reference image that is aligned with the distorted image, first we affine transform it as shown in Fig. 3 left column.

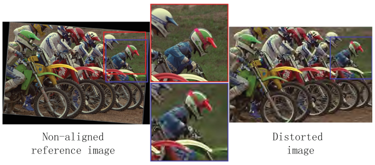


Fig. 3. Nonaligned training samples (Color figure online)

Then from within the border, we randomly sample multiple 224 × 224 sub-images from both the transformed reference image and the distorted image, cantering at the same coordinates. As can be seen from Fig. 3 middle column, the red box and the blue box correspond to one pair of sub-image for training, and the content in the red box and the blue box are similar but not aligned. Finally we collected hundred thousand pairs of samples for training our proposed DCNN model. A stride of 20 has been applied to extract sub images. The same strategy was applied for both the LIVE and the TID2008 datasets, where 80 % of the data was used for training, and the rest for testing. The performance is comparable when sub images are smaller, *e.g.* 32×32. For succinctness, we omit the experiments with smaller sub images. The next subsection reports the overall performance.

4.1 Overall NAR-IQA Performance

LIVE dataset consists of 779 distorted images with one of the following distortion types: JP2k compression(JP2K), JPEG compression(JPEG), White Gaussian(WN), Gaussian blur(BLUR) and Fast Fading(FF) derived from 29 reference images. The subjective evaluations give the Differential Mean Opinion Scores (DMOS) for each of the distorted images. To compare the performance of different IQA methods, we calculated the correlation of the predicted score with the ground-truth DMOS score, and higher correlation indicates better consistency with human assessment, and thus better performance. Specifically, two widely applied correlation criterions were applied in our experiment: Linear Correlation Coefficient (LCC) and Spearman Rank Order Correlation Coefficient (SROCC). LCC reveals the linear dependence between two quantities, and SROCC measures how well the relationship between two variables can be described using a monotonic function.

The results are listed in Table 2, where we further compared with the following benchmarks: FSIM [3], PSNR and SSIM [5] that are FR-IQA methods, and CNN-NR [9], BRISQUE [11] and CORNIA [10] that are NR-IQA methods. We trained and tested the proposed DCNN with the original reference image (FR-DCNN) and also with the affine transformed reference image (NAR-DCNN). As training the CNN is time and storage consuming, we randomly selected 80 % training images of dataset five times. The results of Tables 2 and 3 appeared in the paper is the median evaluation. We found our architecture was rather robust to data splitting.

Table 2. LCC and SROCC score for the LIVE dataset

	FSIM	PSNR	SSIM	BRISQUE	CNN-NR	CORNIA	FR-DCNN	NAR-DCNN	CNN-NR-d
LCC	0.960	0.856	0.906	0.942	0.953	0.935	**0.977**	0.976	0.968
SROCC	0.964	0.866	0.913	0.940	0.956	0.942	**0.975**	**0.975**	0.967

As can be seen from Table 2, when applying our proposed DCNN for FR-IQA, the obtained LCC and SROCC score both achieved best consistency with subjection evaluation. In the NAR-IQA case, our DCNN model outperformed all the listed benchmark methods, and was only slightly worse than the FR-IQA case.

It is important to also study whether the performance gain comes from the use of deep architecture or through the use of a non-aligned reference image. Hence we used one path of DCNN and add *fc#* as well as a regression layer to construct a image quality assessment model. Such model is termed as CNN-NR-d. It applied exactly the same parameters in Table 1. In Tables 2 and 3, we compared the performance of CNN-NR-d to a reported NR-IQA method, the CNN-NR [9] method, as well as our FR-DCNN and NAR-DCNN models. According to the results, we can conclude that, CNN-NR-d did perform better than CNN-NR [9] which shows that DCNN was able to capture more discriminative feature to describe the scene. In addition, the performance of either the FR-DCNN and NAR-DCNN further outperformed CNN-NR-d, which tells that a reference, either aligned or non-aligned, has clearly provided helpful information for IQA.

TID2008 dataset consists of 1700 distorted images derived from 25 reference images. The subjective evaluations give the Mean Opinion Score (MOS) for each of the distorted images. There are 4 types of distortions that are common to the LIVE dataset: JPEG2000, JPEG, WN and GB. These types were considered in our experiments as in many previous works [9,22,23]. Some performances were cited from the published paper, which were tested in a slight different way with an logistic regression. As shown in Table 3, DCNN achieved best consistency with subjection evaluation for the FR-IQA problem, and got comparable performance in the NAR-IQA problem.

Table 3. LCC and SROCC score for the TID2008 dataset

	FSIM	PSNR	SSIM	BRISQUE	CNN-NR	CORNIA	FR-DCNN	NAR-DCNN	CNN-NR-d
LCC	0.926	0.836	0.893	0.892	0.903	0.880	**0.955**	0.941	0.920
SROCC	0.947	0.870	0.902	0.882	0.920	0.890	**0.954**	0.937	0.921

In-house dataset consists of 1050 distorted images derived from 21 high quality images. Each of the 21 image has a reference image that contains similar scene but is not related with the distorted image under any known geometrical transformation. The distorted images deteriorated in the same way that LIVE dataset [24] generated five types of distortion images. Basically, the readme attached with Live dataset was followed. For FF distortion, some adjustments were made as readme of generating FF are difficult to follow. Each distortion type has a same number of distorted images for each reference image on In-house dataset. All the reference and similar reference images were downloaded from Internet with Google or Flickr by searching same keywords, for example "road".

Five of the images are presented in Fig. 1. As can be seen, the scenes between each pair are similar but actually contain different contents. Compared to the previous two datasets that used synthesized non-aligned reference image, the in-house dataset was collected with more realistic setup for the NAR-IQA problem. Specifically, to collect one pair of data, we first collected two high-quality images denoted as I_A and I_B respectively, and then we downgraded I_A to I_A^*. In order to obtain a ground-truth quality score for I_A^*, we used our FR-DCNN method to predict an image quality score for I_A^*, and the score was then regarded as the ground-truth score to evaluate various NAR-IQA and NR-IQA methods. We have also applied the strategy but with the FSIM method to generate another set of ground-truth score. For this strict NAR-IQA setup, we compared with the FSIM method of similar scene reference images, and the results are listed in Table 4. Larger DMOS but smaller FSIM indicates worse image quality, thus the sign of results in Table 4 just indicates anti-correlation or correlation and was ignored. The NAR-DCNN model in Table 2 which were trained with affine transformed image pairs were utilized for experiments of In-house dataset in Tables 4 and 6.

Table 4. LCC and SROCC score for the in-house dataset

GT by FRDCNN	NAR-DCNN	CNN-NR-d	CORNIA	BRISQUE	DIIVINE	NAR-FSIM
LCC	**0.893**	0.880	0.856	0.753	0.737	0.174
SROCC	**0.892**	0.872	0.864	0.756	0.746	0.157
GT by FSIM	NAR-DCNN	CNN-NR-d	CORNIA	BRISQUE	DIIVINE	NAR-FSIM
LCC	0.690	0.684	**0.750**	0.582	0.640	0.234
SROCC	0.835	0.823	**0.907**	0.734	0.754	0.160

It is interesting to see that, first the FSIM algorithm is very sensitive to the mis-alignment between the distorted image and the reference image, and under the NAR-IQA case, FSIM performed very poorly. Second, training on LIVE and testing on new dataset proved great generalization capability of our algorithm.

Third, although NAR-DCNN model were trained on LIVE dataset with random affine transformation, our proposed NAR-DCNN model obtained superior consistency than the benchmark methods, which demonstrates that the presented model can effectively mine the similar scene structural information between the distorted image and the reference image for quality assessment. In fact, as can be seen from Fig. 1, in many cases, although the reference image “looks” similar to the distorted image, they might actually be captured at different locations.

4.2 The Influence of Distortion Type

Which type of distortions could be best modeled in the proposed NAR-IQA setup is an interesting question. Hence we further conducted distortion-specific experiments on LIVE dataset, and the results are listed in Tables 5 and 6. It is clear to see that for most of the stated distortion type, DCNN achieved the best performance under either the FR-IQA problem or the NAR-IQA problem. Interestingly, NAR-DCNN performs less effective for the WN case. We believe it is because that the white noise was added to the image at pixel level. The simulated geometrical transform has spread the white noise, which may break

Table 5. LCC and SROCC score *vs.* Distortion type for the LIVE dataset

LCC	JPEG2k	JPEG	WN	BLUR	FF	SROCC	JPEG2k	JPEG	WN	BLUR	FF
FSIM	0.910	0.985	0.976	0.978	0.912		0.970	**0.981**	0.967	0.972	**0.949**
PSNR	0.873	0.876	0.926	0.779	0.870		0.870	0.885	0.942	0.763	0.874
SSIM	0.921	0.955	0.982	0.893	0.939		0.939	0.946	0.964	0.907	0.941
CNN-NR	0.953	0.981	0.984	0.953	0.933		0.952	0.977	0.978	0.962	0.908
CORNIA	0.951	0.965	**0.987**	0.968	0.917		0.943	0.955	0.976	0.969	0.906
BRISQUE	0.922	0.973	0.985	0.951	0.903		0.914	0.965	**0.979**	0.951	0.877
FR-DCNN	0.972	**0.990**	0.980	**0.990**	**0.975**		0.977	**0.981**	0.950	**0.991**	0.948
NAR-DCNN	**0.981**	0.983	0.964	0.982	0.965		**0.984**	0.976	0.884	0.983	0.918

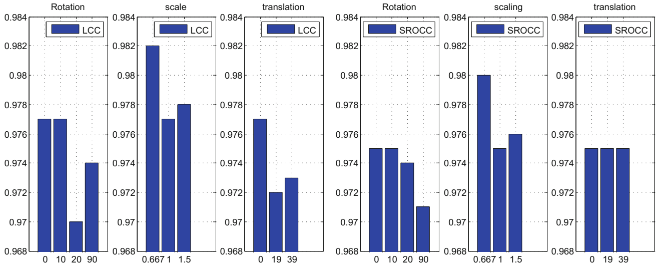
Table 6. LCC and SROCC score *vs.* Distortion type for in-house dataset

LCC	Ground-truth by FR-DCNN					Ground-truth by FSIM				
	JPEG2k	JPEG	WN	BLUR	FF	JPEG2k	JPEG	WN	BLUR	FF
brisque	0.645	0.733	0.648	0.726	0.486	0.613	0.588	0.744	0.563	0.472
cornia	0.801	0.835	0.594	0.861	0.784	**0.852**	**0.828**	0.638	**0.783**	**0.841**
diivine	0.507	0.718	0.691	0.570	0.534	0.438	0.641	**0.807**	0.501	0.493
CNN-NR-d	**0.911**	**0.944**	0.663	0.726	**0.794**	0.796	0.794	0.781	0.546	0.828
NAR-CNN	0.862	0.909	**0.776**	**0.867**	0.746	0.844	0.768	0.593	0.732	0.746
SROCC	Ground-truth by FR-DCNN					Ground-truth by FSIM				
	JPEG2k	JPEG	WN	BLUR	FF	JPEG2k	JPEG	WN	BLUR	FF
brisque	0.674	0.670	0.556	0.681	0.508	0.677	0.714	0.862	0.631	0.510
cornia	0.616	0.665	0.527	**0.847**	0.770	0.768	0.805	0.673	**0.902**	**0.868**
diivine	0.361	0.533	0.611	0.542	0.449	0.228	0.528	**0.919**	0.567	0.370
CNN-NR-d	**0.812**	**0.921**	0.544	0.700	**0.788**	0.604	0.806	0.826	0.657	0.833
NAR-CNN	0.761	0.852	**0.696**	0.833	0.775	**0.787**	**0.836**	0.644	0.799	0.792

down the structural information. Thus, the proposed DCNN did not discover suitable features for this case. The invariance to image content variation has slightly compromised the discriminating power to the pixel aligned distortion in this case.

4.3 The Influence of Structural Similarity

The Influence of Geometrical Transformation. Since we claim that the proposed DCNN is capable to utilize non-aligned reference image, it is important to analyze the influence of geometrical variation to the NAR-IQA process. Hence we applied rotation, scaling and translation transform individually to the reference image for LIVE dataset and compared their performance. Specifically, for rotation we tested the following values $\pi/18$, $\pi/9$ or $\pi/2$, for scaling we tested shrinking to 0.667 or enlarge to 1.5, and for translation we tested one-tenth (19 pixels) and one-fifth (39 pixels) of sub image size. As Fig. 4 shows, our proposed DCNN is stable to translation, and performance dropped slightly for rotation and scaling. Existing FR-IQA methods cannot handle the NAR-IQA problem well, even performances of CW-SSIM [16] are only around 0.14, while our DCNN could achieve consistency of 0.97+ with human assessment, significantly higher than these existing FR-IQA methods.



SROCC	aligned	R10	R20	R90	S0.667	S1.5	T19	T39
CW-SSIM	0.850	0.227	0.171	0.026	0.007	0.025	0.158	0.111
NAR-DCNN	**0.975**	**0.975**	**0.974**	**0.971**	**0.980**	**0.976**	**0.975**	**0.975**

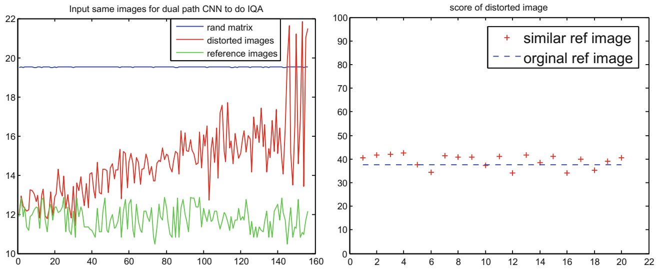
Fig. 4. The influence of geometrical transformation to LCC and SROCC of the proposed DCNN for NAR-IQA (R, S, T in the table means rotation, scale, translation respectively)

The models were retrained in above different situations, with different geometrical variation images. However, different geometrical variation cases can be handled in our mixture variation cases in Tables 2 and 3. As shown in Fig. 4, our NAR-DCNN were very insensitive to affine transform as it reserve the structure well. Next we will discuss what happened if the structural similarity don't exist.

IQA referenced with different structural similarities. Now, the NAR-DCNN model was fed with distortion images and random selected high quality images. The reference images almost have no structural similarity with distortion images. This experiment is to demonstrate influence of structural similarities in similar reference image selection. The performance of our NAR-DCNN dropped heavily to (**LCC:0.932**, **SROCC:0.924**), but amazingly, it still better than some existed methods. We believe our NAR-DCNN has also learnt some non-structural feature, "naturalness". Next, reference image same as distorted images were provided to control structural similarity and to explore the "naturalness".

In Fig. 5(a), we listed results in the following setups: both reference and distorted image using same reference images of LIVE dataset (green line), using same distorted images of LIVE dataset (red line), and using randomized 2D matrix (blue line). Since it uses reference image to guide the assessment of image quality, by giving different reference images, the quality score should change accordingly. If reference image is selected to be same as distorted image, then the predicted quality should be very good, because according to the "reference", the distorted image is "perfect". As "naturalness" features worked for our model, statistically, the image assessment Q would be: $Q(I_r, I_r) > Q(I_d, I_d) > Q(I_m, I_m)$, which the experimental results support well. The x axis is the sequence number of the tested distorted image, the higher DMOS value of the distorted images, the larger sequence number it would be. It shows that our DCNN did extract non-structural features, "naturalness" for quality assessment. The quality of the reference image need to be carefully controlled.

Although using 2 different non-aligned reference images may give different scores, they will correlate to each other very well. In fact, as shown in Fig. 5(b), we demonstrate that applying 20 similar scene reference images to NAR-IQA problem for one distorted image of 'whitehouse'. The blue dash line indicates scores predicted by NAR-DCNN(trained with DMOS) referenced with original aligned image, while red '+' demonstrates situation referenced with different similar scenes. The discrepancy between cases referenced with similar scene image and original image was in the range of $[-5, 5]$. Averaging the scores by more reference images could produce more stable results. Algorithms to retrieve very similar images for IQA was also appealing to be exploited. This part will be explored in the future. Different similar reference images and more predicted results will be depicted in the supplementary material.



(a)

(b)

Fig. 5. Comparison of IQA score with different structural similarities: (a)IQA with reference image same as distorted image(b)IQA with different similar reference images (Color figure online)

The selection of similar reference images do have a influence on the performance of NAR-DCNN model. However, the selection of reference image was rather robust with large geometrical variation and image content changing. NAR-DCNN gave state of art performance when these experiments were referenced with random affine transformed reference images of different similarities, large geometrical variation and even content changing images in the paper. These experiments demonstrate that our NAR-DCNN model is very robust to the selection of non-aligned reference images. Although we haven't work out a accurate measurement for IQA similar reference images selection, we believe the selection of structural similarity is not harsh. NAR-DCNN largely benefits from comparison between the nonaligned reference and the distorted images, especially by capturing the structural information.

5 Conclusion and Future Work

This paper presents a Dual-path deep CNN model for image quality assessment using non-aligned reference images with similar scene. The proposed method validates that, the NAR-IQA problem is solvable where an aligned reference image is not available, but a non-aligned image with similar scene can be well used as reference. The proposed DCNN model handles the IQA problem well, and it is observed that DCNN could use non-aligned reference images and achieve superior quality assessment consistency than many existing methods that use aligned reference images.

The next step of the work includes exploring measurement for IQA similar reference images selection, collecting larger non-aligned IQA dataset and building deep model of different architecture to further improve prediction consistency, as well as applying the technique to certain real-world applications.

Acknowledgment. This work is partially supported by National Science Foundation of China under Grant NO. 61473219, and the National Basic Research Program of China (973 Program) under Grant No. 2015CB351705.

References

1. Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., Battisti, F.: TID 2008-a database for evaluation of full-reference visual quality assessment metrics. Adv. Mod. Radioelectron. **10**(4), 30–45 (2009)
2. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Trans. Image Process. **15**(2), 430–444 (2006)
3. Zhang, L., Zhang, L., Mou, X., Zhang, D.: FSIM: a feature similarity index for image quality assessment. IEEE Trans. Image Process. **20**(8), 2378–2386 (2011)
4. Pei, S.C., Chen, L.H.: Image quality assessment using human visual dog model fused with random forest. IEEE Trans. Image Process. **24**(11), 3282–3292 (2015)
5. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
6. Sheikh, H.R., Bovik, A.C., De Veciana, G.: An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Trans. Image Process. **14**(12), 2117–2128 (2005)
7. Li, Q., Wang, Z.: Reduced-reference image quality assessment using divisive normalization-based image representation. IEEE J. Sel. Top. Sig. Process. **3**(2), 202–211 (2009)
8. Rehman, A., Wang, Z.: Reduced-reference image quality assessment by structural similarity estimation. IEEE Trans. Image Process. **21**(8), 3378–3389 (2012)
9. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1733–1740. IEEE (2014)
10. Ye, P., Kumar, J., Kang, L., Doermann, D.: Unsupervised feature learning framework for no-reference image quality assessment. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1098–1105. IEEE (2012)

11. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process. **21**(12), 4695–4708 (2012)
12. Gao, X., Lu, W., Tao, D., Li, X.: Image quality assessment and human visual system. In: Visual Communications and Image Processing 2010, International Society for Optics and Photonics, pp. 77440Z–77440Z (2010)
13. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2004, vol. 2, pp. 1398–1402. IEEE (2003)
14. Posner, M.I., Petersen, S.E.: The attention system of the human brain. Technical report, DTIC Document (1989)
15. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **11**, 1254–1259 (1998)
16. Sampat, M.P., Wang, Z., Gupta, S., Bovik, A.C., Markey, M.K.: Complex wavelet structural similarity: a new image similarity index. IEEE Trans. Image Process. **18**(11), 2385–2401 (2009)
17. Damera-Venkata, N., Kite, T.D., Geisler, W.S., Evans, B.L., Bovik, A.C.: Image quality assessment based on a degradation model. IEEE Trans. Image Process. **9**(4), 636–650 (2000)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
19. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R.: Signature verification using a time delay neural network. Int. J. Pattern Recogn. Artif. Intell. **7**(04), 669–688 (1993)
20. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 539–546. IEEE (2005)
21. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding (2014). arXiv preprint: arXiv:1408.5093
22. Xue, W., Zhang, L., Mou, X.: Learning without human scores for blind image quality assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 995–1002 (2013)
23. Moorthy, A.K., Bovik, A.C.: Blind image quality assessment: from natural scene statistics to perceptual quality. IEEE Trans. Image Process. **20**(12), 3350–3364 (2011)
24. Sheikh, H.R., Sabir, M.F., Bovik, A.C.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans. Image Process. **15**(11), 3440–3451 (2006)