Multimed Tools Appl (2018) 77:14859–14872
DOI 10.1007/s11042-017-5070-6

Saliency-based deep convolutional neural network for no-reference image quality assessment

Sen Jia[1] · **Yang Zhang**[2]

Received: 1 January 2017 / Revised: 25 July 2017 / Accepted: 31 July 2017 /
Published online: 22 August 2017

Abstract In this paper, we proposed a novel method for No-Reference Image Quality Assessment (NR-IQA) by combining deep Convolutional Neural Network (CNN) with saliency map. We first investigate the effect of depth of CNNs for NR-IQA by comparing our proposed ten-layer Deep CNN (DCNN) for NR-IQA with the state-of-the-art CNN architecture proposed by Kang et al. (2014). Our results show that the DCNN architecture can deliver a higher accuracy on the LIVE dataset. To mimic human vision, we introduce saliency maps combining with CNN to propose a Saliency-based DCNN (SDCNN) framework for NR-IQA. We compute a saliency map for each image and both the map and the image are split into small patches. Each image patch is assigned with a patch importance value based on its saliency patch. A set of Salient Image Patches (SIPs) are selected according to their saliency and we only apply the model on those SIPs to predict the quality score for the whole image. Our experimental results show that the SDCNN framework is superior to other state-of-the-art approaches on the widely used LIVE dataset. The TID2008 and the CISQ image quality datasets are utilised to report cross-dataset results. The results indicate that our proposed SDCNN can generalise well on other datasets.

Keywords NR-IQA · CNN · Saliency map

✉ Sen Jia
sen.jia@bristol.ac.uk

Yang Zhang
yang.zhang@bristol.ac.uk

[1] Intelligent Systems Laboratory, University of Bristol, Bristol, UK

[2] Bristol Vision Institute, University of Bristol, Bristol, UK

1 Introduction

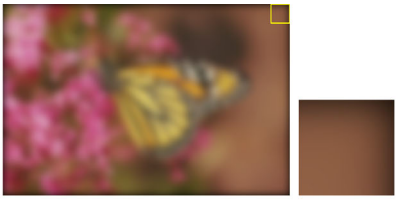
Numerous digital photos are published on social media websites everyday. The image quality is hugely varied due to the conditions they were captured or the data types used for storage. Low quality images may result in a bad user experience, therefore many Image Quality Assessment (IQA) approaches have been proposed to objectively classify image quality. The ground truth of IQA is different from traditional object classification tasks because IQA label is subjectively marked by observers. Each individual assesses image samples based on their own point of view so different scores may be assigned to the same image. When using a reference image, viewers can better rate an image by comparing the distorted one against its undistorted version, so called Full-Reference IQA (FR-IQA). While in many cases, assessment can only be made based on a single distorted photo since the reference image is not available, known as No-Reference IQA (NR-IQA).

In the work of [24], they listed three types of knowledge are essential in building a successful QA model: 1. Distortion type - what causes the distortion, e.g, gaussian white noise or gaussian blur; 2. Image source - how a model can capture discriminant information to distinguish good quality from bad; 3. Human Visual System (HVS) - physiology about how human beings view an image. This paper aims at solving all the three points by combining Convolutional Neural Networks (CNNs) with saliency map.

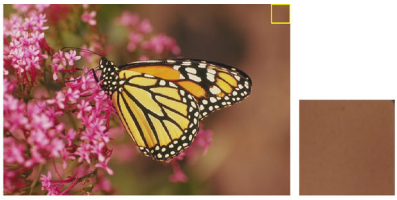
For the first qustion, early IQA methods were designed specifically for one distortion type. For instance, Sheikh et al. [20] proposed a NR-IQA method for JPEG2000 compression by combining gaussian scale mixture and wavelet coefficient. Most recent work tries to solve a more challenging task where an algorithm can be applied on more than one distortion type or the distortion type is unknown. The first question has been well studied, so all the works we compare with can be applied on different distortions, see Table 4.

The second question is actually how to design a good system that can accurately estimate image quality. For NR-IQA, methods can be grouped into two main categories, Natural Scence Statistics (NSS)-based and training-based. The former one aims at seeking "naturalness" among undistorted images so that "unnatural" distortion signal can be easily detected [6, 14–16, 18]. For the training-based method, it relies on a set of features learned from images and then a classifier is trained. The training-based method can be considered as a traditional machine learning task [10, 11, 13, 25, 27]. Therefore how to extract discriminant features is a common question between vision recognition tasks and training-based IQA. It has been shown that CNN-learned features outperform hand-designed ones (local binary patterns or scale-invariant feature transform) in many areas, such as object classification [7], face gender recognition [8] or fashion detection [9]. More recently, CNNs have been introduced to NR-IQA and achieved state-of-the-art results [10, 11, 13]. It has also been shown that the depth of CNNs plays an important role in feature extraction [7, 8, 22, 23]. The CNN architecture used in our work consists of ten convolutional layers, which is deeper than prior works, referred as Deep CNN (DCNN) in this paper.

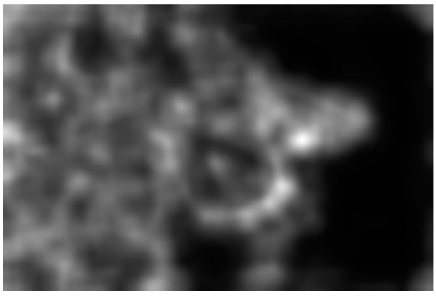
For the third question, the use of saliency map has shown to be helpful in IQA tasks [28]. However, most existing CNN methods have not introduced HVS into IQA designs. In the works of [10, 11], a whole image is split into a set of small patches and the overall estimation is based on the average of those patches. Two similar image patches may be labelled as two totally different noise levels. In Fig. 1, human can easily classify the quality of the two whole images, but it is difficult to recognise the difference based on the two upper-right patches. It may lead to a low assessment accuracy that assigning equal weight to all patches within an image. To introduce HVS, we combine CNNs with the saliency map algorithm of [19], referred as Saliency-based DCNN (SDCNN). One close work to ours is



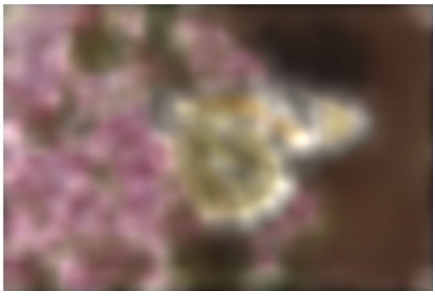
(a) An image with Gaussian Blur and a patch from its upper right corner.



(b) The reference image and a patch from its upper right corner.



(c) The saliency map computed on the image with Gaussian Blur.



(d) Merging the distorted image and its saliency map.

Fig. 1 Two similar patches from two images with totally different holistic quality

that of [13] in which a gradient map is used to measure the importance of each patch, see Section 2.

In this paper, we train a DCNN model on the LIVE dataset [21] for comparability. Following the same experiment setting in [10, 11], the proposed DCNN achieves higher LCC and SROCC scores, 0.9782 and 0.9735 respectively. Our experiment shows saliency maps can further improve CNNs for NR-IQA. Our SDCNN achieves state-of-the-art results on LIVE. To validate the generalisation ability of our SDCNN, we train an SDCNN model on the LIVE dataset and apply it on the TID2008 [17] and the CSIQ [2] datasets for cross-dataset evaluation, see Section 4.4.1.

We begin by reviewing current state-of-the-art techniques in Section 2. Then, in Section 3 we describe our CNN architecture and other pre-processing steps. The proposed method is evaluated on the LIVE, TID2008 and CSIQ datasets in Section 4, while conclusions are drawn in Section 5.

2 Related work

As mentioned above, the NSS-based NR-IQA method tries to capture statistical properties of undistorted images regardless of the content. To compute NSS features, most algorithms firstly transform an image into another domain to formulate distributions or train a model. In the work of [16], the NSS feature is computed on a set of wavelet coeffiecients. Their work needs to identify the distortion type before applying a distortion-specific classifier. Similarly, Saad et al. [18] transform each image using discrete consine transform and the resulting coefficients are used for a generalized gaussian density model. Later,

Li et al. [14] proposed a NR-IQA method using neural network to extract features in the domain of shearlet. But the auto-encoder used in their work is different from CNNs. The former one is designed for unsupervised dimensionality reduction while the latter one has been shown to achieve state-of-the-art results in many vision tasks [7–9, 22, 23]. More recently, Hadizadeh and Bajicb [6] extract a richer quality feature (e.g. gradient magnitues and its Laplacian) from the wavelet domain, so called wavelet-packet. Most NSS-based methods need to transform images into a new domain before extracting features. This transformation may be time consuming or only focus on a specific type of information (wavelet, DCT or shearlet). Mittal et al. [15] proposed an NSS IQA method which is directly applied in the spatial domain. In their work, it has been shown that mean subtracted contrast normalized coefficients can represent statistical properties of distortion after applying the local normalisation. Recent CNN-based NR-IQA methods [10, 11], including ours, are also based on the same spatial domain. But the difference is that we try to use CNNs to learn quality features instead of seeking the naturalness.

On the other hand, much effort has been made recently in converting the IQA task into a machine learning problem. Ye et al. [25] proposed a method that building a codebook for image patches. The trainin process of their work is similar to CNN-based methods that the learned quality feature is not hand-crafted. Later, the idea of codebook was combined with object detection by Zhang et al. [27]. An interesting point is that object detection is actually similar to saliency maps but the saliency map is more suitable for "free-view" tasks [3]. Feng et al. [5] proposed a NR-IQA method based on salient image patches. But it may be a bottleneck for quality estimation that the feature learning step they used (sparse coding). To leverage the power of CNN, Kang et al. [10, 11] proposed a simple CNN architecture on local normalised images using [15]. But the depth of their CNN model may limit the power of feature extraction and assigning eqaul weight to all patches may not consistent with HVS. One close work to ours was proposed by Li et al. [13], which also combines CNNs with a saliency algorithm of gradients. In their work, a two-layer CNN model was used for feature extraction. More imporantly, they segment each image and then apply the Prewitt operator to detect edges. However, weighing on edges may lose the attention on other important factors for image quality, such as contrast sensitivity or luminance [1, 29].

The proposed method tries to leverage the power of DCNN for feature extraction. Following the concept proposed in [22] that combining small perceptive fields with more CNN layers, our proposed DCNN model contains ten convolutional layers. To better mimic HVS for NR-IQA, we use saliency maps to measure the importance of each image patch. In the work of [28], most existing saliency algorithms can improve the accuracy for IQA. We choose the method of [19] for availability.

3 Methodologies

Similar to the work of [10], we apply local normalisation on each image in those datasets. The normalised image is split into Image Patches (IPs) to train a DCNN model. Those image patches are assigned with the same quality label as the input image. The main difference between our architecture and the CNN in [10] is that more CNN layers are used to extract quality feature for NR-IQA. For the proposed SDCNN, as shown in Fig. 2, we compute saliency map for each image and the map is also split into Saliency Patches (SPs). The SP is used to determine whether the associated IP is a Salient IP (SIP) for the trained model to predict on, see Section 3.3.

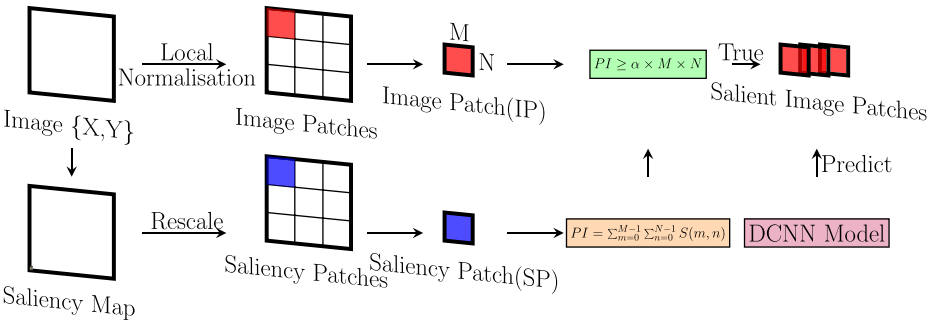


Fig. 2 Proposed SDCNN workflow

3.1 Local normalization

A contrast normalization has been applied before training. We locally normalise contrast on every image in LIVE before training a CNN. Given an input image, we calculate the normalised pixel $\hat{I}(i, j)$ at the location of (i, j) within the window W by:

$$\mu(i, j) = \frac{\sum_{m=-M}^{m=M} \sum_{n=-N}^{n=N} I(i+m, j+n)}{(2M+1)(2N+1)} \tag{1}$$

$$\sigma(i, j) = \frac{\sqrt{\sum_{m=-M}^{m=M} \sum_{n=-N}^{n=N} (I(i+m, j+n) - \mu(I))^2}}{(2M+1)(2N+1)} \tag{2}$$

$$\hat{I}(i, j) = \frac{I(i, j) - \mu(i, j)}{\sigma(i, j) + C} \tag{3}$$

We set C=1 in cases of the divisor is zero, and the size of W is 7×7(M=N=3).

3.2 DCNN architecture

Kang et al. [10] used only one convolutional layer followed by maxpool and minpool layers. In the work of [12], they visualised the convolutional kernels at the first layer that the CNN-learned features are selective to frequency and orientation of an image. In Kang's work, they also showed the CNN-learned features for NR-IQA. But it is difficult to tranlate CNN features into a human-readable format. Extensive works [7, 8, 22, 23] have shown a CNN architecture with more layers can deliver a better featre extraction. Our proposed CNN architecture is similar to [22] that we stack ten convolutional layers with small receptive fields for NR-IQA.

Following the setting used in the work of [10], we split each input image into small patches in the size of 32×32. To build a deep CNN architecture, we introduce the idea proposed in [22] by stacking small kernels (3×3) as an efficient representation of large kernels. A 2×2 maxpool layer is added and the number of kernels is doubled every two convolutional layers. Two fully connected layers are added at the end of the model, each of which has 2048 neurons. Dropout is added in the two fully connected layers with ratio of 0.5. Table 1 illustrates the architecture of our proposed DCNN.

We apply exponential linear units [4] after each convolutional and fully-connected layer.

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ \beta(exp(x) - 1), & \text{if } x \leq 0 \end{cases} \tag{4}$$

Table 1 DCNN architecture

Layer Type	Input	Conv	Conv	Maxpool	Conv	Conv	Maxpool	Conv	Conv	Maxpool	Conv	Conv	Maxpool	Conv	Conv	Fully-Con	Fully-Con	Softmax
Size and Number	32×32 @3	3×3 @32	3×3 @32	2×2	3×3 @64	3×3 @64	2×2	3×3 @128	3×3 @128	2×2	3×3 @256	3×3 @256	2×2	3×3 @512	3×3 @512	2048	2048	–

when x is greater than zero, its output is same as Rectified Linear Units. But when x is less or equal than zero, the function squashes the output to a negative constant value, negative one in our experiment(β=1).

3.3 SDCNN algorithm

The proposed saliency algorithm in [19] (known as SDSR) has been proven to be one of the most stable saliency maps on distorted images [28]. The SDSR is based on similarity between a pixel value and its neighbourhood. As shown in Fig. 1c, the saliency map mimics human attention by focusing on important image regions.

The workflow of our SDCNN is shown in Fig. 2. For each distorted image X, we compute the saliency map using the algorithm of [19]. The resulting saliency values S in the map are in the range of [0,255]. The higher saliency value it has, the more salient image pixel is. We rescale saliency map values into the range of [0,1]. For the ith image patch IP_i, we compute its Patch Importance PI_i by:

$$PI_i = \sum_{m=0}^{m=M-1} \sum_{n=0}^{n=N-1} S(m,n) \tag{5}$$

where $S(m,n)$ is the saliency value at the location of m,n in the SP_i. The PI_i is in the range of [0, $M \times N$].

We set a importance coefficient α as a threshold to select SIPs for quality score prediction. The IP_i is considered to be a SIP if $PI_i \geq \alpha \times M \times N$. In our experiment, the α is chosen from {0,0.01,0.1,0.5}. A higher α value leads to a more salient SIP subset.

When assessing an image, we only apply the trained model on the SIP subset to predict quality scores. The final score of the whole image is the average of those scores computed on SIPs. Note that the SDCNN method applies on all image patches when $\alpha = 0$, which is equivalent to DCNN.

4 Experiments and results

4.1 Datasets

The LIVE [21] dataset is used to train and test our DCNN and SDCNN methods. The TID2008 [17] and the CSIQ [2] datasets are only used for SDCNN cross-dataset evaluation. We train a classifier on the four common types (JPEG, JP2K, WN, GBLUR) from the LIVE dataset and evaluate the model on the same distortion types from the other two datasets.

4.1.1 LIVE IQA dataset

This dataset contains 799 distorted from five types of quality distortion, JPEG, JP2K, White Noise(WN), Gaussian Blur(GBLUR) and Fast Fading(FF). LIVE also comes with 29 reference images. The subjective labels for this dataset are Difference Mean Opinion Score (DMOS) in the range of [0,99] (higher DMOS denotes lower quality). In this work, we randomly separate the whole dataset into training, validation and test sets, then each image is split into smaller non-overlapping patches.

4.1.2 TID2008 IQA dataset

The TID2008 dataset contains 17 different distortion types and each of which includes 100 distorted images. Note that the label for this dataset is Mean Opinion Score (MOS) in the range of [0,9] (higher MOS value indicates better quality). To evaluate the classifier trained on LIVE, a non-linear mapping function is applied to convert the LIVE DMOS label to TID2008 MOS. The whole TID2008 dataset is split to 80% for training and the rest for testing.

4.1.3 CSIQ IQA dataset

The CSIQ dataset is also used for cross-dataset validation. The label of this dataset is DMOS but in the range of [0,1]. To evaluate the model on this dataset, we only rescale the LIVE prediction into the same range of CSIQ. No non-linear mapping function is required.

4.2 Evaluation measurements

We evaluate our model using two measurements, LCC and SROCC. LCC is used to measure the strength of correlation between predictions and ground truth on vaildation and test set. While SROCC is high if ground truth can be monotonically represented by predictions on the same set.

For distortion-specific experiment, all the images from a specific distortion type (e.g. JPEG) are split into 60% for training, 20% for validation and 20% for testing. For non-distortion-specific experiment, all images from the LIVE dataset are split following the same protocol. During each train-test iteration, the best test result is recorded according to the highest LCC obtained on the validation set. We repeat this train-test iteration 10 times to report the average accuracy on the test set.

For the cross-dataset validation, we use the four types of distortions that are shared by the three datasets. For the TID2008 dataset, 80% of the total is used to train a non-linear mapping function and the rest is for testing. The whole CSIQ dataset is used for testing because the mapping procedure is not required. We repeat this cross-dataset evaluation 30 times to report the average accuracy. Note that the saliency coefficient is applied in SDCNN during the test phase, see Section 4.4.

4.3 DCNN experiments

Firstly, the distortion type is known before training a CNN model, so called distortion-specific assessment. Each image in the training set is split into small IPs in the size of 32×32 and we assign the same label as the original image to all those patches. The starting learning rate and momentum are 0.01 and 0.9. The total number of training epochs is 15 and the batch size used is 64. After every five epochs, the learning rate and momentum are reduced by scaling and subtracting 0.1 respectively.

Table 2 illustrates the average LCC and SROCC of ten iterations on the test set. For the distortion-specific experiment, our DCNN outperforms the CNN work [10] on most distortion types, especially on "Fast Fading", 0.9697 LCC and 0.9504 SROCC. But CNN performs better on the type of "JPEG" (see Fig. 3a–e).

For non-distortion-specific assessment, all images from the LIVE dataset are used for training regardless their distortion types, denoted as "ALL" in Table 2. Following the same measurements used in the distortion-specific assessment, higher LCC and SROCC are obtained by proposed DCNN, 0.9782 and 0.9735 respectively (see Fig. 3f).

Table 2 Average DCNN performance on the LIVE test set

Distortion Type	JP2K	JPEG	WN	GBLUR	FF	ALL
CNN [10]-LCC	.9530	**.9810**	.9840	.9530	.9330	.9530
DCNN-LCC	**.9782**	**.9810**	**.9933**	**.9756**	**.9697**	**.9782**
CNN [10]-SROCC	.9520	**.9770**	.9780	.9620	.9080	.9560
DCNN-SROCC	**.9691**	.9532	**.9901**	**.9702**	**.9504**	**.9735**

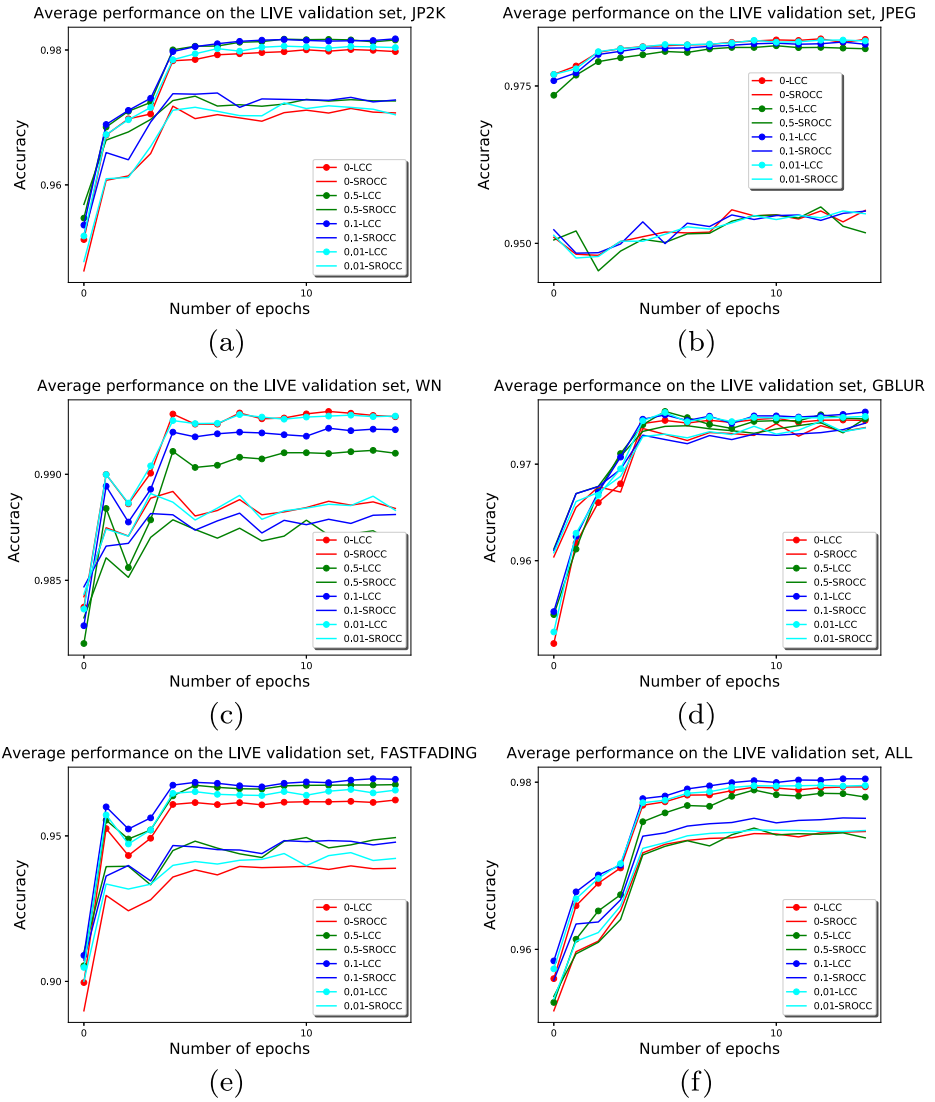
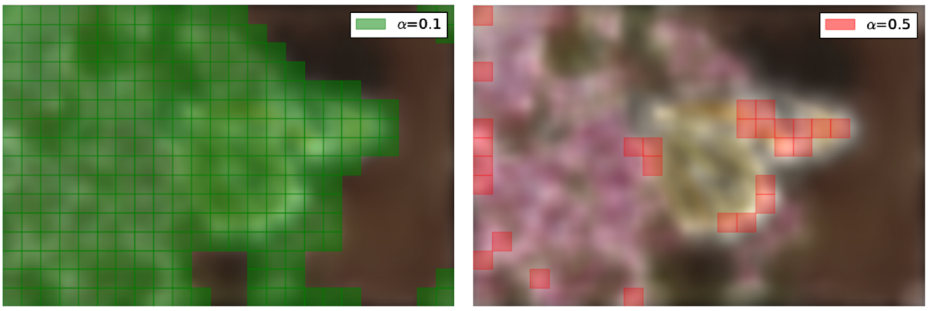


Fig. 3 LCC and SROCC accuracies on the LIVE validation set using different importance coefficient ($\alpha = [0, 0.01, 0.1, 0.5]$)



(a) α=0.1 (b) α=0.5

Fig. 4 The trained model is evaluated on different subsets of patches (**a**), (**b**) based on the importance coefficients α used.

4.4 SDCNN Experiments

The second experiment is to combine the proposed DCNN model with the saliency map of [19]. As discussed in Section 3.3, the importance coefficient α is applied to ignore insignificant image patches when predicting the quality score. In this experiment, the value of $\alpha \in \{0, 0.01, 0.1, 0.5\}$. When $\alpha = 0$, the SDCNN model is the same as DCNN because it considers all IPs as SIPs. When we set the α to a large value (e.g. $\alpha = 0.5$), the SDCNN only predicts on SIP subset, see Fig. 4c. All experiment settings used for the SDCNN are the same as used in last experiment DCNN.

Figure 3 shows the average LCC and SROCC learning curve of ten iterations on the LIVE validation set. Different α values are applied to different distortion types. As shown in Fig. 3a, d–e, saliency map improves accuracy on the three distortion types of "JP2K", "GBLUR" and "FASTFADING". The reason behind could be people focus more on salient regions when marking the three distortion types. Our proposed SDCNN also achieves improvement for non-distortion-specific accuracy on LIVE Fig. 3f.

The average LCC and SROCC of ten iterations on the LIVE validation set are used to choose the best importance coefficient α^* for each distortion type, see Table 3. For "JPEG" and "WN", the highest LCC on the validation set is obtained when $\alpha = 0$. That is the average performance of SDCNN on the test set is the same as DCNN on the two distortion types because no saliency map is applied during the test phase, as shown in Table 4.

The performance of SDCNN on the LIVE test set based using chosen α^* is reported to compare against other methods in Table 4, the best results are highlighted in bold. Our proposed SDCNN outperforms other IQA methods on most distortion types. Especially on the "ALL" distortion type, the proposed method achieves 0.9794 LCC and 0.9757 SROCC, which outperforms the state-of-the-art FR-IQA method in [26]. The method in [6] achieved

Table 3 The highest average LCC on the LIVE validation set and the best importance coefficient α^*

Distortion Type	JP2K	JPEG	WN	GBLUR	FF	ALL
α^*	0.1	0	0	0.5	0.1	0.1
LCC	.9829	.9832	.9934	.9766	.9708	.9807

Table 4 LCC and SROCC on the LIVE test set

LCC	JP2K	JPEG	WN	GBLUR	FF	ALL
DIVINE [16]	.9220	.9210	.9880	.9230	.8880	.9170
BLIIDNS-II [18]	.9348	.9676	.9799	.9381	.8955	.9302
BRISQUE [15]	.9229	.9734	.9851	.9506	.9030	.9424
CORNIA [25]	.9510	.9650	.9870	.9680	.9170	.9350
SOM [27]	.9520	.9610	.9910	.9740	.9540	.9620
SESANIA [14]	.9537	.9732	.9806	.9749	.9195	.9476
Hadizade [6]	.9591	.9720	**.9954**	.9717	.9345	.9601
Prewitt [13]	.9780	.9770	.9930	.9450	.9600	.9660
SFOSR [5]	.9390	.9510	.9490	.9450	.9250	.9640
CNN [10]	.9530	**.9810**	.9840	.9530	.9330	.9530
DCNN	.9782	**.9810**	.9933	**.9756**	.9697	.9782
SDCNN	**.9801**	**.9810**	.9933	.9744	**.9723**	**.9794**
SROCC	JP2K	JPEG	WN	GBLUR	FF	ALL
DIVINE [16]	.9130	.9100	.9840	.9210	.8630	.9160
BLIIDNS-II [18]	.9285	.9422	.9691	.9231	.8893	.9306
BRISQUE [15]	.9139	.9647	.9786	.9511	.8768	.9395
CORNIA [25]	.9430	.9550	.9760	.9690	.9060	.9420
SOM [27]	.9470	.9520	.9840	**.9760**	.9370	.9640
SESANIA [14]	.8862	.9293	.9309	.9410	.8807	.9340
Hadizade [6]	.9467	.9768	**.9903**	.9695	.9202	.9521
Prewitt [13]	.9640	.9350	.9880	.9410	.9450	.9580
SFOSR [5]	.9320	.9470	.9820	.9510	.9460	.9530
CNN [10]	.9520	**.9770**	.9780	.9620	.9080	.9560
DCNN	**.9691**	.9532	.9901	.9702	.9504	.9735
SDCNN	**.9691**	.9532	.9901	.9732	**.9513**	**.9757**

the highest result on the type of "WN". The reason behind can be that the first and the second order information can better represent the white noise.

4.4.1 SDCNN cross-dataset test

Deeper CNN architectures normally consist of more parameters which may lead to over-fitting on a small training set. To investigate generalisation ability of our SDCNN, we train a model on the LIVE dataset and test on TID2008 and CSIQ. Only the four common distortion types ("JP2K", "JPEG", "WN", "GBLUR") are used for this cross-dataset experiment. The output label of our model is in the DMOS range of [0,99]. Following the settings in [11], we split the TID2008 dataset into two subsets, 80% of the data is used to train a non-linear mapping and the rest is for testing. We repeat this split 30 times to report cross-dataset performance on the TID2008 dataset. We do not apply non-linear mapping on the CSIQ dataset so that the whole dataset is used to report test result. The best importance coefficient

Table 5 LCC and SROCC on the TID2008 and the CSIQ datasets, trained on the LIVE dataset

LCC	TID2008	CSIQ
CORNIA [25]	.8900	.9140
SFOSR [5]	–	0.7290
CNN[10]	**.9030**	.9130
IQA-CNN++ [11]	.8950	.9280
DCNN	.8865	.9292
SDCNN	.8872	**.9295**
SROCC	TID2008	CSIQ
CORNIA [25]	.8800	.8990
SFOSR [5]	–	0.7400
CNN [10]	**.9200**	.9230
IQA-CNN++ [11]	.9060	**.9360**
DCNN	.8824	.9340
SDCNN	.8853	.9341

$\alpha^* = 0.1$ for SDCNN is chosen based on the non-distortion-specific setting in the last experiment (Table 3). Our results are compared against other state-of-the-art methods in Table 5, the best results are highlighted in bold.

5 Conclusions

In this paper, we have proposed a deep CNN architecture for NR-IQA and achieved state-of-the-art results on different datasets. Comparing with coding-based NR-IQA methods [5, 25], CNN-learned features can deliver a higher performance and better generalise on unseen data. Our method outperforms the work of [10] by leveraging a deeper CNN architecture for quality feature extraction. The saliency map has been shown that can further improve the accuracy of CNNs. The training time for each batch is 0.048 seconds and the total training time for 15 epoches on the whole LIVE dataset is about two hours using NVIDIA GTX 980. It takes the DCNN model 0.042 seconds when evaluating on an image. But computing saliency map is time consuming for evaluation which[19] costs about three seconds for an image from LIVE (average height: 548, average width: 665). An even deeper CNN architecture (hundreds of layers [7]) might deliver a better performance but existing quality datasets are much smaller than the ones for object recognition. Merging quality dataset is also very difficult due to the subjectiveness of annotation. Nevertheless, CNN-learned feature offeres a new promising way for NR-IQA.

Acknowledgements The authors would like to thank Dr. Thomas Lansdall-Welfare for the help in English wrting.

References

1. Barten P (1999) Contrast sensitivity of the human eye and its effects on image quality. Press Monograph Series. SPIE Optical Engineering Press. https://books.google.co.uk/books?id=RVZRAAAAMAAJ
2. Chandler DM (2010) Most apparent distortion: full-reference image quality assessment and the role of strategy. J Electron Imaging 19(1):011,006. https://doi.org/10.1117/1.3267105
3. Chang KY, Liu TL, Chen HT, Lai SH (2011) Fusing generic objectness and visual saliency for salient object detection. In: IEEE international conference on computer vision (ICCV)
4. Clevert D, Unterthiner T, Hochreiter S (2015) Fast and accurate deep network learning by exponential linear units (ELUs). arXiv:1511.07289
5. Feng T, Deng D, Yan J, Zhang W, Shi W, Zou L (2016) Sparse representation of salient regions for no-reference image quality assessment. Int J Adv Robot Syst 13(5):1729881416669,486. https://doi.org/10.1177/1729881416669486
6. Hadizadeha H, Bajicb IV (2016) No-reference image quality assessment using statistical wavelet-packet features. Pattern Recogn Lett (Accepted)
7. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. arXiv:1512.03385
8. Jia S, Lansdall-Welfare T, Cristianini N (2016) Gender classification by deep learning on millions of weakly labelled images. In: IEEE 16th international conference on data mining workshops (ICDMW), pp 462–467. https://doi.org/10.1109/ICDMW.2016.0072
9. Jia S, Lansdall-Welfare T, Cristianini N (2016) Time series analysis of garment distributions via street webcam. In: International conference image analysis and recognition. Springer, pp 765–773
10. Kang L, Ye P, Li Y, Doermann D (2014) Convolutional neural networks for no-reference image quality assessment. In: IEEE conference on computer vision and pattern recognition, pp 1733–1740. https://doi.org/10.1109/CVPR.2014.224
11. Kang L, Ye P, Li Y, Doermann D (2015) Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In: IEEE international conference on image processing (ICIP), pp 2791–2795. https://doi.org/10.1109/ICIP.2015.7351311
12. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
13. Li J, Zou L, Yan J, Deng D, Qu T, Xie G (2016) No-reference image quality assessment using prewitt magnitude based on convolutional neural networks. SIViP 10(4):609–616. https://doi.org/10.1007/s11760-015-0784-2
14. Li Y, Po LM, Xu X, Feng L, Yuan F, Cheung CH, Cheung KW (2015) No-reference image quality assessment with shearlet transform and deep neural networks. Neurocomputing 154:94–109. https://doi.org/10.1016/j.neucom.2014.12.015. http://www.sciencedirect.com/science/article/pii/S0925231214016798
15. Mittal A, Moorthy AK, Bovik AC (2012) No-reference image quality assessment in the spatial domain. IEEE Trans Image Process 21(12):4695–4708
16. Moorthy AK, Bovik AC (2011) Blind image quality assessment: from natural scene statistics to perceptual quality. IEEE Trans Image Process 20(12):3350–3364. https://doi.org/10.1109/TIP.2011.2147325
17. Ponomarenko N, Lukin V, Zelensky A, Egiazarian K, Carli M, Battisti F (2009) TID2008 - A database for evaluation of full-reference visual quality assessment metrics. Adv Mod Radioelectron 10(4):30–45
18. Saad MA, Bovik AC, Charrier C (2012) Blind image quality assessment: a natural scene statistics approach in the DCT domain. IEEE Trans Image Process 21(8):3339–3352. https://doi.org/10.1109/TIP.2012.2191563
19. Seo HJ, Milanfar P (2009) Static and space-time visual saliency detection by self-resemblance. J Vis 9(12):15. https://doi.org/10.1167/9.12.15
20. Sheikh HR, Bovik AC, Cormack L (2005) No-reference quality assessment using natural scene statistics: JPEG2000. IEEE Trans Image Process 14(11):1918–1927. https://doi.org/10.1109/TIP.2005.854492
21. Sheikh HR, Sabir MF, Bovik AC (2006) A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans Image Process 15(11):3440–3451. https://doi.org/10.1109/TIP.2006.881959
22. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. Eprint Arxiv
23. Szegedy C, Liu W, Jia Y, Sermanet P, Reed SE, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2014) Going deeper with convolutions. arXiv:1409.4842
24. Wang Z, Bovik AC (2011) Reduced- and no-reference image quality assessment. IEEE Signal Proc Mag 28(6):29–40. https://doi.org/10.1109/MSP.2011.942471

25. Ye P, Kumar J, Kang L, Doermann D (2012) Unsupervised feature learning framework for no-reference image quality assessment. In: IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 1098–1105
26. Zhang L, Zhang L, Mou X, Zhang D (2011) FSIM: A feature similarity index for image quality assessment. IEEE Trans Image Process 20(8):2378–2386. https://doi.org/10.1109/TIP.2011.2109730
27. Zhang P, Zhou W, Wu L, Li H (2015) SOM: Semantic Obviousness metric for image quality assessment. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 2394–2402, https://doi.org/10.1109/CVPR.2015.7298853
28. Zhang W, Borji A, Wang Z, Callet PL, Liu H (2016) The application of visual saliency models in objective image quality assessment: a statistical evaluation. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2015.2461603
29. Zhang Y, Agrafiotis D, Naccari M, Mrak M, Bull DR (2013) Visual masking phenomena with high dynamic range content. In: IEEE international conference on image processing (ICIP). IEEE, pp 2284–2288

Sen Jia is a Ph.D candidate and Research Associate at the University of Bristol under the supervision of Nello Cristianini. His current research involves the use of deep learning for computer vision tasks. He is also interested in incremental sub-space based algorithms for scalable data and robustness analysis of noisy labels. He received his M.Sc. degree with Distinction from Newcastle University and his B.E. from Beijing University of Technology.

Yang Zhang received his B.Eng (2008) in a joint training program between Beijing Institute of Technology, China, and the University of Central Lancashire, the M.Sc (2010) and the PhD (2015) from the University of Bristol, respectively. He has worked as an intern researcher at BBC Research and Development, on HEVC range extension standardization for HDR video coding. He is currently working as a research associate in the Visual Information Laboratory in Bristol Vision Institute (BVI), and Microelectronics Group, University of Bristol. His research interests include high dynamic range imaging, perceptual image and video coding, human vision models, objective and subjective video quality assessment, wireless video streaming, machine learning and vibration signal processing.