Hierarchical Feature Degradation based Blind Image Quality Assessment

Jinjian Wu, Jichen Zeng, Yongxu Liu, Guangming Shi
Xidian University, China

Weisi Lin
Nanyang Technological University, Singapore

Abstract

Though blind image quality assessment (BIQA) is highly demanded for many image processing systems, it is extremely difficult for BIQA to accurately predict the quality without the guide of the reference image. In this paper, we introduce a novel BIQA method with hierarchical feature degradation (HFD). Since the human brain presents hierarchical procedure for visual recognition, we suggest that different levels of distortion generate different degradations on hierarchical features, and propose to consider the degradations on both the low and high level features for quality assessment. Inspired by the orientation selectivity (OS) mechanism in the primary visual cortex, an OS based local visual structure is designed for low-level visual content extraction. Meanwhile, according to the feature integration function of deep neural networks, the deep semantics is extracted with the residual network for high-level visual content representation. Next, by analyzing the degradation on both the local structure and the deep semantics, a HFD based memory (prior knowledge) is learned to represent the generalized quality degradation. Finally, with the guidance of the HFD based memory, a novel HFD-BIQA model is built. Experimental results on the publicly available databases demonstrate the quality prediction accuracy of the proposed HFD-BIQA, and verify that the HFD-BIQA performs highly consistent with the subjective perception[1].

1. Introduction

With the tremendously increase of digital photographs in our daily life, it is highly desired to faithfully evaluate the qualities in many signal processing systems, e.g., digital signal acquisition, compression, transmission, and so on. Thus, a reliable objective image quality assessment (IQA) method, which performs consistently with the subjective perception, is greatly demanded.

During the past decade, a large amount of IQA methods have been proposed. The largest number of these IQA methods are full-reference and reduced-reference, which require the whole reference image or part of the reference information for quality prediction. However, the reference information is always unavailable for most situations. Thus, the no-reference/blind IQA (BIQA), which do not need any reference information for IQA, is required, and BIQA has became a popular research topic in the recent years. In this work, we focus on designing a novel BIQA method.

Since we know nothing about the reference image, it is extremely difficult for BIQA to accurately evaluate the qualities of images. At the early state, most BIQA methods usually employed the prior knowledge of the distortion type for quality prediction, which are called distortion-specified BIQA, e.g., sharpness for blur [1], blockiness for JPEG [2], ringing for JPEG2000 [3], and so on. These distortion-specified BIQAs only work for their corresponding types of distortion, and have a limited application scope.

Recently, researches focus on the non-distortion-specific BIQA methods [4]. The most popular type of BIQA methods are the natural scene statistical (NSS) based, which follow the assumption that the low-level features if nature scenes present some kind of statistical distributions, and the distortions will degrade such distributions. For example, Moorthy et al. [5] suggested to learn the NSS with the generalized Gaussian distribution (GGD) in the wavelet domain, and measured the image quality as the change on the GGD coefficients (i.e., DIIVINE method). Following this work, Saad et al. analyzed the NSS characteristic on the DCT coefficients of images, and introduced a BLIINDS method in [6]. Moreover, Mittal et al. extended the NSS work in the spatial domain, and introduced the BRISQUE in [7]. Recently, Zhang et al. further extended the NSS work and integrated a large set of NSS features in several domains

[1]This work was supported by the National Natural Science Foundation of China (Nos. 61401325, 61472301, 61632019, 61621005).

BIQA (i.e., the IL-NIQE method) [8]. Besides these NSS methods, Ye and Doermann [9] proposed to measure the quality with the guide of a trained codebook, and Zhang et al. [10] introduced to learn a local quantized pattern based visual codebook to guide the BIQA. These low-level feature based BIQA methods still have a large gap from the human perception.

In order to design a novel BIQA method which performs more consistently with the human perception, we tune to investigate the visual perception of the HVS. Researches on the neuroscience indicate that the HVS can be modeled as a hierarchy for visual feature extraction and visual recognition, i.e., from low-level feature (e.g., local structure) to comprehensive high-level feature (e.g., semantics information) [11, 12]. Inspired by this, we suggest that different levels of distortion generate different degradations on hierarchical features, and propose to consider the degradations on both the low and high features for quality assessment.

The HVS presents obvious orientation selectivity (OS) mechanism in the primary visual cortex for low-level feature (e.g., local edge) extraction [13–15]. Inspired by the OS mechanism, the OS based local structure is introduced for low-level feature extraction. Meanwhile, with multiple layers to learn hierarchical representations of the visual input, the later layers of deep neural network can efficiently extract the high-level features of visual contents [16]. As one of the most powerful deeper network architectures, the residual network (ResNet) [17] is adopted for high-level feature extraction. Next, the degradations on both low and high features are analyzed for memory creation. By learning the correlation between the subjective quality (e.g., the mean opinion score (MOS) from subjective IQA test) and the degraded features with support vector regression (SVR), the prior knowledge database (i.e., memory) about quality degradation is created. Finally, under the guidance of the degradation memory, a novel hierarchical feature degradation (HFD) based BIQA (i.e., HFD-BIQA) method is proposed. With both the low and high level features for quality prediction, the proposed HFD-BIQA outperforms the existing BIQA methods (experimental results have demonstrated that the proposed HFD-BIQA has a remarkable improvement against these existing methods).

2. Hierarchical Quality Degradation Measurement

In this section, the hierarchical subjective perception on visual quality degradation is firstly briefly analyzed. Then, inspired by the OS mechanism, the OS based local structure is introduced for low-level feature extraction. Next, according to the ResNet, the high-level feature from the latest layer is extracted for deep semantics representation. Finally, the degradation on both low and high level features are analyzed to create the HFD based memory for BIQA.

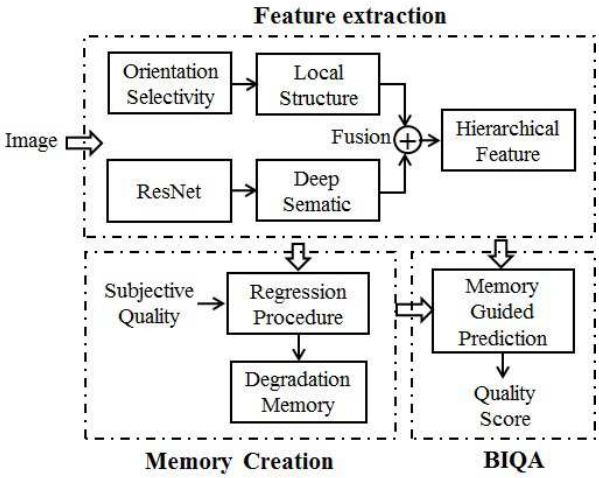

Figure 1: The architecture of the HFD-BIQA method.

The architecture of the proposed BIQA model is shown in Fig. 1.

2.1. Hierarchical Quality Degradation

Researches on cognitive neuroscience indicate that the HVS is a hierarchy of cortical areas, in which the input visual signal is hierarchically processed with increasingly sophisticated representation (from low to high level features) [11, 12]. For the input visual image, the primary visual areas (V1 and V2) are adapted to extract simply local features. e.g., edge and orientation. The successive areas (V3, V4, and medial-temporal area) integrate these local features and generalize some global representations, e.g., contour and shape. By further integrating the output of these global representations, the high-level visual areas (inferotemporal and prefrontal areas) finally generate the high-level semantics, e.g., abstract and categories.

Different levels of noise will generate hierarchical distortions on visual contents, and cause individual degradations on visual qualities. As shown in Fig. 2, the *Hats* image is distorted by three different levels of Gaussian blur noise (WBN). A weak noise level (PSNR=$36.71dB$) in Fig. 2 (a) has slightly blurred the local edge, while has little effect on the shape of the hats. With the increase of noise level, the local edge in Fig. 2 (b) (with PSNR=$26.37dB$) is obviously distorted, while the concept can still be clearly extracted (i.e., precisely understanding this is an image with three hats). Under a strong noise level (PSNR=$20.93dB$), the local edge and shape in Fig. 2 (c) is seriously distorted, and it is impossible to extract the accurate concept information (hats or air balloon or something else) from this image. Due to the hierarchical perception model in the HVS, we need consider the degradations on multi-levels of features (e.g., low and high levels of features) for BIQA mod-

(a) PSNR=36.71dB (b) PSNR=26.37dB (c) PSNR=20.93dB

Figure 2: Hierarchical visual quality degradation under different noise levels.

eling.

2.2. Local Visual Structure Extraction

It is well known that the HVS is highly adapted to extract the local structure for visual perception, and thus the structural degradation is widely used for quality assessment [18, 19]. Researches on neuroscience demonstrate that neurons on the primary visual cortex present substantial OS for low-level structure extraction [13]. Moreover, the OS arises from the intracortical responses (i.e., excitatory and inhibitory interactions) among neurons in a local receptive field [20]. Inspired by the OS mechanism, we suggest to represent the local structure ($\mathcal{F}_l$) with both response intensity ($\mathcal{I}_r$) and response pattern ($\mathcal{P}_r$) in a local region.

The HVS is extremely sensitive to luminance changes, and the response intensity is directly related to the change/difference on luminance. Thus, for a given image I, the intensity of the local structure for each pixel can be represented as its luminance change, and is calculated as the gradient magnitude,

$$\mathcal{I}_r(x) = \sqrt{(G_h(x))^2 + (G_v(x))^2}, \tag{1}$$

where G_h and G_v are the horizontal and vertical gradient maps, which can be acquired with Prewitt filters.

As an invariant feature of image, visual patterns have been widely used in many visual recognition work [21, 22]. The response pattern $\mathcal{P}_r$ of a local receptive field is decided by the arrangement of intracortical responses (i.e., excitatory and inhibitory interactions). Moreover, neighbor neurons with similar preferred orientations always present excitatory interactions, and these dissimilar ones present inhibitory interactions [23]. Thus, the pattern form $\mathcal{P}_r(x)$ that each pixel presented is described as the arrangement of interactions between the central pixel x and pixels in its local neighborhood ($\mathcal{R}(x)$=$\{x_1, x_2, \cdots, x_n\}$) [24],

$$\mathcal{P}_r(x) = \mathcal{A}(\mathcal{I}(x|x_1), \mathcal{I}(x|x_2), \cdots, \mathcal{I}(x|x_n)), \tag{2}$$

where $\mathcal{I}(x|x_i)$ is the interaction type between x and x_i,

$$\mathcal{I}(x|x_i) = \begin{cases} 1 & \text{if } |\theta(x) - \theta(x_i)| < \mathcal{T} \\ 0 & \text{else} \end{cases}, \tag{3}$$

$$\theta(x) = \arctan \frac{G_v(x)}{G_h(x)}, \tag{4}$$

where '1' represents excitatory interaction, and '0' represents inhibitory interaction. The judging threshold $\mathcal{T}$ determines the interaction type, and is set as $\mathcal{T}$=6° according to the subjective viewing test on visual masking [24].

With Eq. (2) we can see that the number of pattern type is growing exponentially with the pixel number in $\mathcal{R}(x)$ (2^n different types). For example, a 5×5 local region possesses more than 10 million (2^{24}) different types of pattern form, which is too many for efficient visual pattern representation. With further analysis, we have found that not all of these patterns are equally appeared (some types of patterns are more frequently appeared, e.g., patterns which represent smooth and edge regions). Moreover, some patterns have similar format and represent homogeneous visual contents. Therefore, we can select these representative patterns for visual structure representation. And in this work, the K-Means clustering algorithm is adopted for representative pattern selection,

$$\{\hat{\mathcal{P}}_r^k, k = 1, 2, \cdots, K\} = \arg\min \sum_{k=1}^{K} \sum_{m=1}^{M} ||w_m \cdot (\mathcal{P}_r^m - \hat{\mathcal{P}}_r^k)||^2, \tag{5}$$

where K is the number of representative patterns, $\hat{\mathcal{P}}_r^k$ represents the nth clustering centroid, and we set K=1000 in this work (to make sure that the numbers of the low and high level features are the same). w_m is the weight factor and is computed as the appearance probability of $\mathcal{P}_r^m$.

With Eqs. (1) and (5), the response intensity ($\mathcal{I}_r$) and response pattern ($\hat{\mathcal{P}}_r$) for each pixel are calculated for its local structure representation. And the low-level visual con-

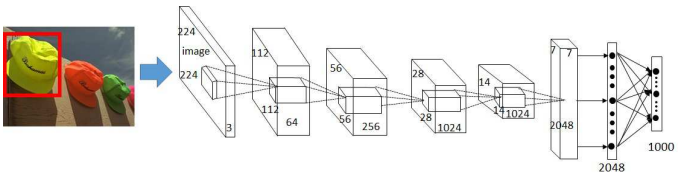

Figure 3: Architecture of the 50-layer ResNet for deep semantics extraction (the latest layer with 1×1000 features).

tent of an image can be mapped into a structure based histogram ($\mathcal{F}_l$),

$$\mathcal{F}_l(k) = \sum_{x=1}^{N} \mathcal{I}_r(x) \cdot \delta(\hat{\mathcal{P}}_r^x, \hat{\mathcal{P}}_r^k) \tag{6}$$

$$\delta(\hat{\mathcal{P}}_r^x, \hat{\mathcal{P}}_r^k) = \begin{cases} 1 & \text{if } \hat{\mathcal{P}}_r^x = \hat{\mathcal{P}}_r^k \\ 0 & \text{else} \end{cases}, \tag{7}$$

where N is the number of pixels in an image.

2.3. Deep Visual Semantics Extraction

As the highest visual area of the HVS, the inferotemporal cortex (IT) integrates the former outputs and generate the high-level visual information (e.g., abstract) for objective recognition [25]. Thus, the high-level visual information plays a key role in visual perception, and distortion on it will severely degrade the quality of image.

Deep learning network can efficiently extract high-level feature for visual recognition. With the inspiration of the hierarchy in the HVS for visual perception, deep learning network uses multiple processing layers to learn and integrate representations, and assemble high-level feature (i.e., deep semantics) in the later layers [16]. Moreover, with the increase of stacked layer number (i.e., the depth of the network), more complex and enrich semantic information can be acquired in the later layers. Therefore, we try to adopt the deep learning network for deep semantics extraction.

As a powerful and deeper neural network, the pretrained ResNet [17] is adopted for deep semantics extraction in this work. Considering the efficiency and computational complexity, a 50-layer ResNet is chosen, whose architecture is shown in Fig. 3. And the output of the latest layer (with 1×1000 features) is used as the deep semantics information (i.e., $\mathcal{F}_h \in R^{1\times1000}$).

2.4. Blind Quality Assessment

With the help of the prior knowledge/memory, the human can accurately predict the quality of an input image. Inspired by this, we try to create a memory database about the hierarchical feature degradation to guide the BIQA. The low/high level features are firstly normalized for fusion,

$$\hat{\mathcal{F}}_i(j) = \frac{\mathcal{F}_i(j)}{\sqrt{\sum_n (\mathcal{F}_i(n))^2}}, \tag{8}$$

where $\mathcal{F}_i$ represents the local structure ($\mathcal{F}_l$) or the deep semantics ($\mathcal{F}_h$), and $\hat{\mathcal{F}}_i(n)$ is the n-th normalized feature.

Next, the two types of features are combined and the hierarchical feature set (i.e., $\mathcal{F}=\{\hat{\mathcal{F}}_l, \hat{\mathcal{F}}_h\}$) is acquire for quality degradation analysis. The correlations between the hierarchical feature sets ($\mathcal{F}$) and the subjective quality scores ($\mathcal{Q}$, i.e., MOS or DMOS) of distorted images are analyzed for degradation memory creation. As an efficient regression procedure from a high dimension to a lower one, the classical support vector regression (SVR) is adopted to learn the mapping relationship between $\mathcal{F}$ and $\mathcal{Q}$. In this work, the LIBSVM [26] is used to acquire the degradation memory ($\mathcal{M}_d$),

$$\mathcal{M}_d = \text{SVR}_{\text{learn}}(\mathcal{F}, \mathcal{Q}). \tag{9}$$

With the guidance of the degradation memory, the quality of an image I can be predicted with its hierarchical feature degradation,

$$\hat{\mathcal{Q}}(I) = \text{SVR}_{\text{predict}}(\mathcal{F}(I), \mathcal{M}_d), \tag{10}$$

where $\mathcal{F}(I)$ is the hierarchical feature set of the input image I, and $\hat{\mathcal{Q}}$ is the predicted quality score

3. Experimental Result Analysis

In this section, the efficiency of the hierarchical feature degradation is firstly illustrated. Then, the prediction accuracy of the HFD-BIQA method is demonstrated by comparing with the existing state-of-the-art BIQA methods on the public available databases.

Three large-scale public IQA databases are chosen for experimental result analysis, which are CSIQ [27], LIVE [28], and TID2013 [29]. The CSIQ database consists of 30 original scenes, and each scene is degraded by 6 types of distortions under 5 different noise levels. The LIVE database contains 29 original scenes, and each scene is degraded by 5 types of distortions. The TID2013 possesses 25 original scenes, and each is degraded by 24 types of distortions under 5 levels.

Meanwhile, three classical metrics for the IQA performance measurement are adopted, which are the Spearman rank order correlation coefficient (SRCC), the Pearson linear correlation coefficient (PLCC), and the root mean squared error (RMSE). The SRCC represents the prediction monotonicity, and a better IQA method returns a larger SRCC value. The PLCC measures the prediction accuracy (the higher PLCC the better performance), and the RMSE represents the prediction deviation (the smaller PMSE the better performance).

3.1. Analysis on Hierarchical Degradation

The HVS hierarchically processes the input visual content, and different levels of distortion generate different

(a) Lady-Face with weak noise

(b) Lady-Face with strong noise

(c) Green-House with weak noise

(d) Green-House with strong noise

Figure 4: An example of hierarchical degradation on two different scenes distorted by JPEG noise under two different levels.

degradations on the hierarchical visual features. An example is shown in Fig. 4, in which two different scenes (i.e., Lady-Face and Green-House from TID2013 [29]) are distorted by JPEG noise under different levels, and the corresponding index values are listed in Tab. 1.

Weak noise mainly degrade the local structure, and has limited influence on the deep semantics. As shown in Fig. 4 (a) and (c), the two images are distorted by weak JPEG noise (with PSNR 28.23 dB and 28.68 dB). As can be seen, though there are obvious degradations on the local structures (e.g., the facial contour in Fig. 4 (a) and the edge of barriers in Fig. 4 (c)), we can still easily extract the primary visual contents of the two images for understanding (i.e., can still understand that Fig. 4 (a) contains a lady face, and Fig. 4 (b) is a green house). Meanwhile, the measurement with local structure can accurately represent the perceptual qualities of the two images. As listed in Tab. 1, Fig.4 (a) (with MOS=3.26) has worse subjective perceptual quality (smaller MOS value) than that of Fig.4 (c) (with MOS=4.64). And the measurement results with local structure is 3.27 and 4.64 for them, which are consistent with the subjective perception (MOS). However, the deep semantics returns an opposite result for the two images (3.61 and 3.20 for them, which means Fig.4 (a) has better quality than Fig.4 (b)).

Strong noise severely degrade the local structure, and directly destroy the deep semantics. As a result, the quality mainly relates to the degradation on the deep semantics, and has little relationship with the degradation on the local structure. As shown in Fig. 4 (b) and (d), the two images are distorted by strong JPEG noise (with PSNR 22.88 dB and 21.61 dB). As a result, we can hardly extract complete information from the two images, e.g., the nose from Fig. 4 (b) or the roof from Fig. 4 (c). Though the local structure is severely distorted, its distortion degree cannot represent the perceptual quality anymore. As shown in Tab. 1, the mea-

Table 1: An example of hierarchical degradation on two different scenes

Fig.4	(a)	(c)	(b)	(d)
MOS	3.26	4.86	2.19	1.66
PSNR	28.23	28.68	22.88	21.61
Local Structure	3.27	4.64	2.41	2.53
Deep Semantics	3.61	3.20	2.11	1.92
HFD-BIQA	3.61	3.63	2.35	1.87

Table 2: Comprehensive analysis of hierarchical degradation on the CSIQ Database

Crit.	Local Structure	Deep Semantics	HFD-BIQA
PLCC	0.847	0.832	0.890
SRCC	0.790	0.762	0.842
RMSE	0.136	0.147	0.120

surement from the local structure returns an opposite result (Fig. 4 (b) has worse quality (2.41) than Fig. 4 (d) (2.53)) against the subjective perception (the MOS for Fig. 4 (b) and (d) are 2.19 and 1.66, respectively). The quality predictions on the two with the deep semantics shows that Fig. 4 (b) (with 2.11) has better quality than that of Fig. 4 (d) (with 1.92), which is consistent with the subjective perception.

The proposed HFD-BIQA can accurately represent the quality degradations on the four images in Fig. 4. By fusing both the low and high features for quality prediction, the proposed HFD-BIQA contains a hierarchical degradation measurement, which can efficiently measure the quality degradation by weak or strong noise. As shown in Tab. 1, the predicted quality for Fig. 4 (a)-(d) are 3.61, 3.63, 2.35, and 1.87, respective. The prediction results shows that Fig. 4 (c) has the best quality, Fig. 4 (a) is the second best, and Fig. 4 (d) is the worst one, which is consistent with the subjective perception.

In order to give a comprehensive analysis on hierarchical feature degradation, the performances of the local structure, the deep semantics, and the proposed HFD-BIQA on the whole CSIQ database [27] are compared, and the comparison results are listed in Tab. 2. By fusing the local structure and the deep semantics, the proposed HFD-BIQA has the highest PLCC and SRCC values, and the lowest RMSE value, which demonstrate that the measurement on the hierarchical feature degradation is more consistent with the subjective perception than that on only one type of feature (i.e., the local structure or the deep semantics).

3.2. IQA Performance Comparison

In order to demonstrate the performance, the proposed HFD-BIQA is compared with 7 state-of-the-art BIQA methods (i.e., IMNSS [10], IL-NIQE [8], NIQE [30], BRISQUE [7], and DIIVINE [5]) and two classical FR-IQA methods (PSNR and MS-SSIM [18]) on the three large IQA databases.

Firstly, the performance of these IQA methods on the individual distortion type of LIVE database is illustrated. There are five different distortion types in LIVE database, namely, JPEG compression noise (JPG), JPEG2000 compression noise (J2K), white Gaussian noise (WGN), Gaussian blur noise (GBN), and fastfading noise (FFN). When building the hierarchical structure degradation memory for the proposed HFD-BIQA, a training procedure is required in the regression module. Similar to the training procedure in these existing BIQA methods (e.g., in [4, 10]), we randomly divide the images that a database contained into two subsets (training and testing subsets). To make sure that there is no overlap between the two subsets, 80% original scenes are randomly selected, and their corresponding distorted images are used for training; the left 20% distorted images are used for testing. Moreover, in order to eliminate the performance bias (not governed by a specific training result), the 80% training - 20% testing procedure is repeated for 1000 times. The median performance across the 1000 times is calculated as the final result.

The performances of these IQA methods on each distortion type of LIVE database are listed in Tab. 3. It is apparent that the HFD-BIQA performs highly consistent with the subjective perception (the SRCC value is larger than 0.9 in all of these distortion types). More concretely, the HFD-BIQA performs the best on two types of distortion (i.e., J2K and FFN) among these BIQA methods, and performs slightly worse than the best one on the other three types. Moreover, by comparing with the FR-IQA methods, the HFD-BIQA outperforms the benchmark PSNR on all of the five distortion types, and performs almost similar with the classic MS-SSIM.

Besides the performance on individual distortion type, the overall performance on the whole database is further analyzed. The overall performances of these IQA methods on the LIVE, CSIQ, and TID2013 databases are listed in Tab. 4. As can be seen, the prediction accuracy of the HFD-BIQA is completely higher than the other BIQA methods (with higher SRCC and PLCC values, and lower RMSE values). Especially for the TID2013 (the largest IQA database which contains 24 different types of distortion, and the existing IQA methods usually perform no good enough on it), the HFD-BIQA achieves a remarkable improvement against these existing BIQA (the SRCC of the HFD-BIQA VS. the second best: 0.764 – 0.643).

Moreover, though BIQA methods are usually hardly

Table 3: Performances (SRCC) comparison on individual distortion type of LIVE database

Distion	NR							FR	
	HFD-BIQA	IMNSS	IL-NIQE	NIQE	BRISQUE	CBIQ	DIIVINE	PSNR	MS-SSIM
J2K	**0.943**	0.934	0.905	0.914	0.914	0.903	0.937	0.904	0.940
JPG	0.951	0.933	0.950	0.937	**0.965**	0.942	0.910	0.891	0.949
WGN	0.972	**0.986**	0.980	0.967	0.979	0.932	0.984	0.984	0.966
GBN	0.919	0.949	0.923	0.931	**0.951**	0.935	0.921	0.797	0.913
FFN	**0.905**	0.8946	0.851	0.861	0.877	0.856	0.863	0.902	0.942

Table 4: Performance Comparison on the whole database (LIVE, CSIQ and TID2013)

DB	Crit.	NR							FR	
		HFD-BIQA	IMNSS	IL-NIQE	NIQE	BRISQUE	CORNIA	DIIVINE	PSNR	MS-SSIM
LIVE	PLCC	**0.951**	0.943	0.905	0.908	0.929	0.937	0.892	0.872	0.945
	SRCC	**0.948**	0.944	0.902	0.908	0.920	0.938	0.882	0.876	0.948
	RMSE	**8.437**	8.705	11.622	11.423	10.421	9.645	12.330	13.360	8.950
CSIQ	PLCC	**0.890**	0.835	0.863	0.726	0.812	0.750	0.804	0.751	0.876
	SRCC	**0.842**	0.789	0.822	0.629	0.748	0.676	0.776	0.806	0.861
	RMSE	**0.120**	0.142	0.130	0.179	0.154	0.172	0.154	0.173	0.133
TID2013	PLCC	**0.764**	0.598	0.641	0.421	0.626	0.552	0.643	0.678	0.790
	SRCC	**0.681**	0.522	0.518	0.330	0.571	0.434	0.567	0.586	0.742
	RMSE	**0.797**	0.997	0.955	1.130	0.931	1.035	0.952	0.911	0.761

matchable with FR-IQA (with the help of the whole reference information, and returns high prediction accuracy), the HFD-BIQA performs better than the benchmark PSNR on all of these databases, and is comparable with the classic MS-SSIM (performs slightly better on LIVE and CSIQ, and a little worse on TID2013).

4. Conclusion

In this paper, we have introduced a novel BIQA method based on hierarchical feature degradation. With the inspiration of hierarchical visual signal processing in the HVS, we have suggested that different levels of distortion generate different degradations on hierarchical features. For example, weak distortion mainly degrades the low-level feature (local structure), and strong distortion directly destroys the high-level feature (deep semantics). Therefore, we proposed to considering both the degradations on the local structure and the deep semantics for quality assessment.

According to the orientation selectivity mechanism, the local visual structure has been extracted for low-level visual content representation. At the meantime, by using the ResNet, the deep semantics has been extracted for high-level visual content representation. By analyzing the degradation on both the local structure and the deep semantics, a hierarchical feature degradation based memory has been learned to guide the BIQA, and the novel HFD-BIQA has been proposed. Experimental results on the large IQA databases have demonstrated that the proposed HFD-BIQA performs highly consistent with the subjective perception.

References

[1] L. Li, W. Xia, W. Lin, Y. Fang, and S. Wang, "No-reference and robust image sharpness evaluation based on multiscale spatial and spectral features," *IEEE Transactions on Multimedia (2017)*, vol. 19, no. 5, pp. 1030–1040, 2017.

[2] F. Pan, X. Lin, S. Rahardja, W. Lin, E. Ong, S. Yao, Z. Lu, and X. Yang, "A locally adaptive algorithm for measuring blocking artifacts in images and videos," *Signal Processing: Image Communication*, vol. 19, no. 6, pp. 499–506, Jul. 2004.

[3] H. Liu, N. Klomp, and I. Heynderickx, "A no-reference metric for perceived ringing artifacts in images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 4, pp. 529–539, Apr. 2010.

[4] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 50–63, Jan. 2015.

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

[6] M. Saad, A. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339 –3352, Aug. 2012.

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

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

[9] P. Ye and D. Doermann, "No-reference image quality assessment using visual codebooks," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3129 –3138, Jul. 2012.

[10] X. Xie, Y. Zhang, J. Wu, G. Shi, and W. Dong, "Bag-of-words feature representation for blind image quality assessment with local quantized pattern," *Neurocomputing*, p. In Press, 2017.

[11] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, pp. 1019–1025, 1999.

[12] S. Hochstein and M. Ahissar, "View from the top: Hierarchies and reverse hierarchies in the visual system," *Neuron*, vol. 36, no. 5, pp. 791–804, Dec. 2002.

[13] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[14] R. Ben Yishai, R. L. Bar-Or, and H. Sompolinsky, "Theory of orientation tuning in visual cortex," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 9, pp. 3844–3848, 1995.

[15] J. Wu, W. Lin, G. Shi, L. Li, and Y. Fang, "Orientation selectivity based visual pattern for reduced-reference image quality assessment," *Information Sciences*, vol. 351, pp. 18–29, Jul. 2016.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[19] J. Wu, W. Lin, G. Shi, and A. Liu, "Reduced-reference image quality assessment with visual information fidelity," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1700–1705, Nov. 2013.

[20] T. W. Troyer, A. E. Krukowski, N. J. Priebe, and K. D. Miller, "Contrast-invariant orientation tuning in cat visual cortex: Thalamocortical input tuning and correlation-based intracortical connectivity," *The Journal of Neuroscience*, vol. 18, no. 15, pp. 5908–5927, 1998.

[21] P. M., H. A., Z. G., and A. T., *Computer Vision Using Local Binary Patterns.* Springer, 207 p, 2011. [Online]. Available: http://www.springer.com/mathematics/book/978-0-85729-747-1

[22] J. Wu, L. Li, W. Dong, G. Shi, W. Lin, and C. C. J. Kuo, "Enhanced just noticeable difference model for images with pattern complexity," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2682–2693, Jun. 2017.

[23] J. A. Cardin, L. A. Palmer, and D. Contreras, "Stimulus feature selectivity in excitatory and inhibitory neurons in primary visual cortex," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 27, no. 39, pp. 333–344, 2007.

[24] J. Wu, W. Lin, G. Shi, Y. Zhang, W. Dong, and Z. Chen, "Visual orientation selectivity based structure description," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4602–4613, Nov. 2015.

[25] L. G. Ungerleider and J. V. Haxby, "'what' and 'where' in the human brain," *Current Opinion in Neurobiology*, vol. 4, no. 2, pp. 157–165, Jan. 1994.

[26] C. C. Chang and C. J. Lin. (2001) Libsvm: a library for support vector machines. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[27] E. C. Larson and D. M. Chandler. (2004) Categorical image quality (csiq) database.

[28] H. R. Sheikh, K. Seshadrinathan, A. K. Moorthy, Z. Wang, A. C. Bovik, and L. K. Cormack. (2006) Image and video quality assessment research at live.

[29] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Kuo, "Color image database TID2013: Peculiarities and preliminary results," in *2013 4th European Workshop on Visual Information Processing (EUVIP)*, Jun. 2013, pp. 106–111.

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