SOM: Semantic Obviousness Metric for Image Quality Assessment

Peng Zhang, Wengang Zhou, Lei Wu, Houqiang Li

Department of Electronic Engineering and Information Science, University of Science and Technology of China
pzhangoo@mail.ustc.edu.cn, zhwg@ustc.edu.cn, wuleibig@gmail.com, lihq@ustc.edu.cn

Abstract

Image quality assessment (IQA) tries to estimate human perception based image visual quality in an objective manner. Existing approaches target this problem with or without reference images. For no-reference image quality assessment, there is no given reference image or any knowledge of the distortion type of the image. Previous approaches measure the image quality from signal level rather than semantic analysis. They typically depend on various features to represent local characteristic of an image.

In this paper we propose a new no-reference (NR) image quality assessment (IQA) framework based on semantic obviousness. We discover that semantic-level factors affect human perception of image quality. With such observation, we explore semantic obviousness as a metric to perceive objects of an image. We propose to extract two types of features, one to measure the semantic obviousness of the image and the other to discover local characteristic. Then the two kinds of features are combined for image quality estimation. The principles proposed in our approach can also be incorporated with many existing IQA algorithms to boost their performance. We evaluate our approach on the LIVE dataset. Our approach is demonstrated to be superior to the existing NR-IQA algorithms and comparable to the state-of-the-art full-reference IQA (FR-IQA) methods. Cross-dataset experiments show the generalization ability of our approach.

1. Introduction

This paper proposes a framework based on semantic obviousness to predict the quality of distorted images to match human perception of image quality. This work targets at the non-distortion-specific NR-IQA problem, which is thought as the most challenging IQA problem since neither knowledge of the distortion type nor the reference image is available. Image quality is a fundamental metric in many computer vision and image processing applications. For example, image quality can be used to measure video quality [25] and to guide an image compression scheme [6].

According to the existence of non-distorted reference images, IQA algorithms can be classified into three categories: full-reference IQA (FR-IQA), no-reference IQA (NR-IQA) and reduced-reference IQA (RR-IQA).

FR-IQA methods [20, 29, 26, 24] require the corresponding non-distorted reference image to predict visual quality of a distorted image. Most of these algorithms estimate visual image quality based on the difference between distorted image and the corresponding reference image. Recent FR-IQA methods usually adopt a top-down framework [11, 29, 24], trying to model the function of human visual system based on some global assumptions on it. Examples of state-of-the-art FR-IQA algorithms include FSIM [29], VIF [20], VSNR [3]. With reference images, the image quality scores produced by FR-IQA methods have high correlation with human perceptual quality.

NR-IQA measures try to estimate the human perceptual quality by extracting discriminative features from distorted images. Most of the traditional NR-IQA algorithms [16, 15, 19, 3, 29, 22, 8] design features based on Natural Scene Statistics (NSS). NSS based approaches process images with certain type of filters, then the responses are used to extract features. Some typical transforms and filters include DCT transform [15, 19], wavelet transform [16, 3] and Gabor filter [29]. NR-IQA can be further classified into two categories: distortion-specific NR-IQA and non-distortion-specific NR-IQA.

Existing IQA methods measure the image quality from signal level rather than semantic sense. FR-IQA methods like [20, 29], NR-IQA algorithms like [15, 3], NSS-based approaches like [15, 29] and training-based methods like [28, 28], even the CNN architecture in [10] all depend on features based on pixel-level or patch-level local characteristic. However, the analysis of eye tracking data [9] shows that most of the time human focus on object-like regions when looking at an image. Therefore, some object-level factors should be explored for image quality assess-

978-1-4673-6964-0/15/$31.00 ©2015 IEEE

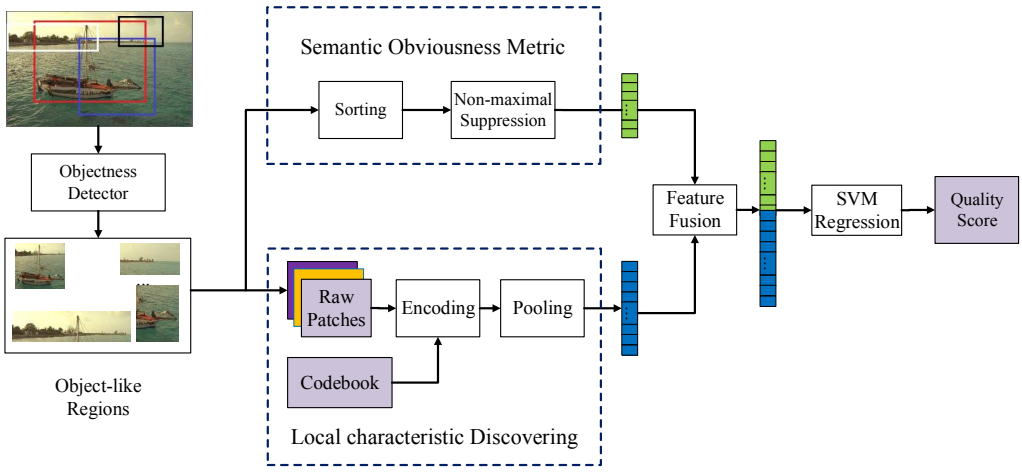


Figure 1. Framework of image quality assessment with semantic obviousness metric .

ment, and that's the fundamental idea of our approach.

In this paper we propose an semantic obviousness based framework to measure image quality. We define semantic obviousness as a metric of how easily can we perceive objects of an image. Any type of distortion may affect the semantic obviousness of an image. We use both semantic obviousness feature and local feature in our approach. First, we design a discriminative feature to represent semantic obviousness of an image. We extract object-like regions from an image and then obtain a semantic obviousness descriptor based on these regions. Second, only the high-scored object-like regions are used to extract features to measure pixel-level local characteristic. Local features are obtained by a codebook based method. We randomly select raw patches to generate local features. Unlike previous methods which use saliency response to weight different image regions, the patches get an automatic weight in our method. The reason is that object-like regions overlap with each other, so that patches from overlapped regions are more likely to be chosen. Besides, we propose a simple yet effective method to increase the number of training samples when the labeled images of IQA dataset is too few to get the model well trained. It is notable that all these principles we mentioned above are algorithm-independent, which means they can be readily incorporated into many existing IQA methods to improve the performance.

The remainder of this paper is organized as follows. In Section 2, we review the related work on NR-IQA and generic object detection. Section 3 introduces our framework in details. Section 4 discusses the experiment details and the analysis of results and parameters. We conclude this paper in the last section.

2. Related Work

Our method to extract local feature is based on some previous work and the proposed semantic obviousness is related to generic object detection.

2.1. No-reference Image Quality Assessment

NR-IQA methods used to depend on hand-crafted features. Most of general NR-IQA methods depend on filtering techniques and various transformations like Gabor filtering, DCT transform, wavelet transform to extract features, which are not originally designed for IQA. Recent advances in machine learning such as convolutional neural network, have revealed the potential to learn discriminative features from raw patches. Ye and Doermann [27] first used codebook in NR-IQA algorithm. They use Gabor-filter based local feature to represent image and use the average of quality scores of codewords to obtain image quality. Although promising performance is achieved, it relies on a subset of labeled data to construct the codebook. Codebook based NR-IQA methods are also used in CORNIA [28]. In that work, the codebook is constructed with unlabeled data and features are learned directly from normalized raw image patches without any filters. Soft assignment and max-pooling are also adopted in this work. Kang *et al.* [10] proposed a new convolutional neural network architecture for NR-IQA. Their network extracts discriminative features from 32×32 patches with a single convolutional layer and a pooling layer, and then estimates image quality score of each patch. They average the scores of all the 32×32 patches to obtain a quality estimation for the whole image. The works above reveal that it's possible to obtain discriminative features directly from raw image patches and machine learning techniques like soft encoding and pooling are useful for IQA frameworks.

2.2. Generic Object Detection

Inspired by human visual system which can perceive objects before identifying them, researchers [2, 1, 7, 4] try to design an objectness detector which is generic over categories. These objectness measures are typically applied to reduce the number of patches that classifiers need to process in a sliding window fashion. Cheng *et al*. [4] proposed an extremely fast generic object detector: BING. It is designed to detect generic objects with well-defined boundary. The computation of BING consists of several simple atomic operations (*e.g*. ADD, BITWISE SHIFT, *et al*.), so it can run on a single machine at a speed of 300 images per second.

Generic object detection is closely related to bottom up attention (saliency), which is used by some full-reference IQA methods [23, 13]. Tong *et al*. [23] proposed a NR-IQA algorithm based on saliency map analysis. In their method, the contribution of any region to the global image quality score of an image is weighted by a function of its saliency. [17] shows that the improvement of IQA algorithms is not guaranteed if saliency response is simply used as a weighting term.

3. Semantic Obviousness based Framework for NR-IQA

Our image quality estimation framework is illustrated in Figure 1. All the object-like regions are extracted by an objectness detector [4]. From these regions we extract two kinds of features: one to measure the global semantic obviousness of the image and one to discover local characteristic. After that, the global and local features are fused to be more discriminative. The final image quality score is obtained by a regression SVM.

3.1. Object-like Region Detection

Computation efficiency is an important factor when applying NR-IQA to other computer vision applications. So we choose the objectness detector BING [4], which is extremely fast and shows high object detection rate and good generalization ability. The work shows that generic objects with well-defined closed boundaries share similar appearance when looking at the norm of the gradients, after resizing their corresponding image patches to small fixed size (*e.g*. 8×8 in BING). Therefore, BING resizes image windows to 8×8 and uses the norm of gradients as a 64D feature to learn a generic objectness measure. The top-scored object-like regions of a test image is shown in Figure 2. As shown in Figure 2, not all the extracted regions are actual objects, but they are distinct from the neighbor regions so that they can attract more attention.

3.2. Feature Extraction

Semantic Obviousness Feature Extraction: For each image in the dataset, we extract all its object-like regions, each with a detection score. The detection score is a metric of the region's possibility of being an object. Distorted regions are more likely to get lower detection scores compared to their corresponding non-distorted regions. Typically, the objectness detector yields more than 3,000 object-like regions. These regions are sorted in descending order based on the corresponding detection scores. We use Intersection over Union (IoU) to measure the overlapping ratio between two regions. IoU is defined as follow:

$$IoU = \frac{S_a \cap S_b}{S_a \cup S_b} \tag{1}$$

where S_a denotes the area of region a, S_b denotes the area of region b.

To reduce the redundancy of these regions, Non-Maximal Suppression (NMS) is performed. A region is removed if its IoU is larger than a threshold α with respect to a region with greater score. The semantic obviousness feature $X = [x_1, x_2, \ldots, x_K]^T$ consists of the detection scores of the top K object-like regions, where x_i is the detection score of the i^{th} region. Distribution of these detection scores contains information about the semantic obviousness. For example, the mean of K scores can basically reflect an image's distortion degree (see experimental results in Section 4.3).

Local Feature Extraction: Unlike previous approaches [14, 11, 15], we extract local descriptors from the object-like regions instead of the whole image. Since an image contains a large number of object-like regions and they typically overlap each other. Only the N top-scored regions are selected using the NMS strategy mentioned above. We denote the set of these regions as S. Then M different $B \times B$ raw image patches are uniformly and randomly sampled from S. Each patch is normalized by the mean and deviation of its elements. For each patch, we concatenate its columns to generate a vector as its descriptor. As a result, for S, we obtain a local descriptor $Y = [y_1, y_2, \ldots, y_M]$, where $y_i \in R^d$, $d = B \times B$. To reduce correlation between features, we apply ZCA whitening [5] on the local descriptor Y.

Since the sampling process is in a random manner, image patches from overlapped regions are more likely to be chosen. That means different patches are weighted automatically based on their probability of belonging to an object region.

Training Sample Augmentation: An important problem associated with the training based IQA algorithms is the small number of training images. For example, Ye *et al*. [28] trained a codebook based NR-IQA model on LIVE IQA dataset. Their quality score is the output of a SVM

Figure 2. Images (on the left) and extracted object-like regions (on the right)

whose input is a feature of dimension 20000×1. LIVE dataset contains only 779 distorted images of 5 distortion types. For distortion-specific experiment, 80% images of the distortion type are used as training set, namely only about 128 training samples for each distortion type. As a result, it may suffer the lack of enough training samples, which is likely to incur over-fitting problem.

To address this problem, we extract E ($E > 1$) features from a training image by simply repeat the feature extraction process E times. Let S_N be the number of patches in extracted region set S of which each feature uses only M randomly selected patches. In our experimental setting, S_N is $40+$ times larger than M (*e.g.* $S_N > 3,000,000, M = 50,000$), so the overlap between E features is negligible.

3.3. Local Feature Encoding

Previous works [28, 27] have shown that codebook based approach can learn efficient features for image quality measuring. We follow this idea and encode the extracted local feature to a codebook constructed on an unlabeled dataset.

Codebook Construction: We followed the codebook construction method introduced in [28], and construct our codebook on an unlabeled dataset. For all distorted images of the dataset, $B \times B$ raw patches are extracted and then normalized and whitened. These patches are then clustered by the K-means algorithm. The constructed codebook is a matrix $D_{d \times W} = [d_1, d_2, \ldots, d_W]$, where $d_i (d_i \in R^d, d = B \times B)$ are normalized cluster centroids.

The database used to construct codebook is CSIQ IQA dataset [12]. It consists of 30 reference images and each is distorted by 6 types of distortions at 4 to 5 degradation levels. There are six types of distortions: JPEG, JP2k, global contrast decrements, additive pink Gaussian noise and Gaussian blurring. The reference images of CSIQ has no overlap with LIVE or TID2008.

Encoding: Local features are quantized by performing soft-assignment coding on the codebook D. The similarity between the i^{th} local feature y_i and the j^{th} codeword d_j is computed by their dot-product as: $s_{i,j} = <y_i, d_j>$. Local feature y_i is encoded as follows [28]:

$$\begin{aligned} c_i = &[max(s_{i,0}, 0), \ldots, max(s_{i,W}, 0), \\ &max(-s_{i,1}, 0), \ldots, max(-s_{i,W}, 0)]^T \end{aligned} \tag{2}$$

3.4. Feature Pooling

In the encoding step, we get a coefficient matrix $C_{2W \times M} = [c_1, c_2, \ldots, c_M]$, where $c_i = [c_{i,1}, c_{i,2}, \ldots, c_{i,2W}]^T$ is obtained by Equation(2). We convert the matrix to a vector with pooling technique for the use of our regression model. Pooling is usually used to fuse features of same type from different regions of an image in machine learning algorithms (*e.g.* CNN). It can efficiently cut down the length of feature thus to prevent over-fitting and reduce amount of computation. Besides, the feature generated by pooling is translation invariant since patch details are hidden. Max-pooling is the most commonly used method. We perform max-pooling on each row of the coefficient matrix. After pooling, we get a feature in the form of:

$$Z = [z_1, z_2, \ldots, z_{2W}]^T \tag{3}$$

where z_i is the maximum of the i^{th} row in coefficient matrix $C_{2W \times M}$. The final feature Z represents the local characteristic of the image.

3.5. Feature Fusion

After all these steps, we get two features for an input image: X measures the semantic obviousness and Z represents the local characteristic. Semantic obviousness is a metric of how easily can human perceive objects of an image. It measures image quality on a semantic level and is basically consistent with human perception of image quality (see analysis in 4.3). Some local distortions may not have significant influence on semantic obviousness, so we also adopt local features to obtain this kind of information. The combination of the two kinds of features enables our approach to measure image quality both on semantic and pixel level. The final descriptor F is in the form as follows:

$$F_{(K+2W)\times 1} = [x_1, \ldots, x_K, z_1, \ldots, z_{2W}]^T \qquad (4)$$

3.6. Regression

After sorting, NMS, encoding and pooling, non-linearities are introduced to the F. Linear models can be applied to non-linear features to obtain promising results in many applications, such as CNN (the output layer), classification algorithms [5] and IQA [27]. We use support vector machine (SVM) regression on the feature F with a linear kernel. Non-linear kernels (RBF) are also tested, and there is no significant improvement in performance. So linear SVM is adopted for efficiency. The output of regression is the final quality score.

4. Experiment

In this section, we first introduce the experimental setting in Section 4.1. Then, in Section 4.2 we study the impact of key parameters. Experimental results on LIVE dataset are shown in Section 4.3. In Section 4.4 cross dataset evaluation is performed. We incorporate our methods with other IQA methods and show the performance in Section 4.5;

4.1. Experimental Settings

To evaluate the proposed algorithm, we conduct experiments on the following two IQA benchmark datasets.

LIVE [21]: This is the most commonly used dataset for the evaluation of IQA algorithms. LIVE consists of 29 reference images with 779 distorted images. There are five different types of distortions: JPEG2000 (JP2K), JPEG, white noise (WN), Gaussian blurring (BLUR) and fast fading Rayleigh channel distortion (FF). A different mean opinion score (DMOS) is provided for each distorted image. DMOS is in the range $[0, 100]$, and a higher DMOS stands for poorer visual quality.

TID2008 [18]: It consists of 25 reference images with 1700 distorted images from 17 different distortion types at 4 degradation levels. Some of the reference images are from the LIVE database. Four distortion types of this dataset are shared with LIVE: JPEG, JP2k, WN and BLUR. A mean opinion score is provided for each image. MOS is in the range $[0, 9]$, and a higher MOS indicates higher visual quality.

Evaluation Method: Following the previous methods, we use two metrics to evaluate the performance of the proposed method: Linear Correlation Coefficient (LCC) and Spearman Rank Order Correlation Coefficient (SROCC). LCC is used to measure the linear dependence between true scores and the predicted quality scores. SROCC measures the strength of association between the predicted scores and true scores according to their monotonic relationship. Given n distorted images, the human perceptual scores V and the predicted scores P, V_i and P_i are converted to their ranks v_i and p_i, and SROCC value is computed from:

$$SROCC = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad (5)$$

where $d_i = v_i - p_i$ is the difference between ranks.

The experimental results we listed are obtained by 100 train-test iterations, and in each iteration 80% of the reference images and their distorted versions are randomly selected as training set and the remaining 20% as testing set.

4.2. Impact of Parameters

In the proposed method, some parameters are important for the performance. In this section, we'll discuss the impact of those parameters. We'll focus on some parameters related to generic object detection and semantic obviousness feature extraction. Size of raw patches is fixed at 7×7, codebook size is set to 10000 and IoU threshold α is fixed to 0.8 when perform Non-Maximal Suppression to remove overlapped object-like regions. The parameter E is set to 3 since we find best performance is obtained when 3 local features are obtained from one image for training. By default, when we focus on one parameter, all the other parameters are set to their optimal values. All the results in this section are obtained on LIVE dataset.

Dimension of Semantic Obviousness Feature: Our semantic obviousness feature $X = [x_1, \ldots, x_K]$ is composed of detection scores of the top K object-like regions. To choose the optimal value for K, we extract 2000 object-like regions for all images, and show the variances of detection scores in Figure 3. As shown in the figure, when $K > 1000$, the variance of detection scores is almost zero. In other words, detection scores of regions out of the top 1000 are less informative. Besides, Cheng *et al*. [4] has shown that 1000 regions generated by BING can cover nearly all (96.2%) potential object regions. Due to these reasons, we set K to 1000 in all our experiments.

Number of Object-like Regions for Local Feature Extraction: In the proposed approach, we only use patches from N top-scored regions to extract local features. To

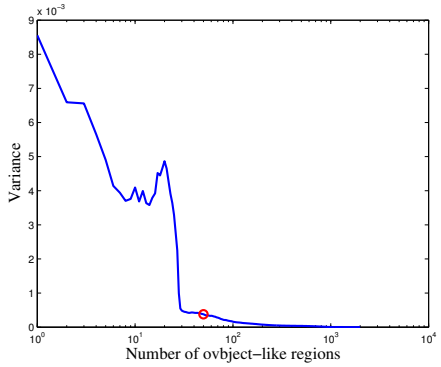


Figure 3. The variance of detection scores of all images on LIVE

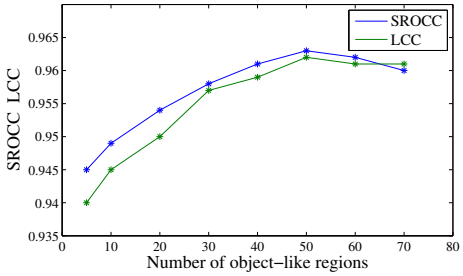


Figure 4. SROCC and LCC with respect to the number of object-like regions

get the proper value of N, we fix the other parameters and test the algorithm with N set to 10,20,...,70. As shown in Figure 4, the best performance is obtained when about 50 object-like regions are extracted. When N is too small, the performance is poor because most area of the image remains unused. On the other hand, if N is too large, some object regions get over-weighted, which will cause a decrease in performance too.

Number of Patches for Local Feature Extraction In our feature extraction, M raw patches are randomly sampled from the N object-like regions to obtain a feature to represent the local characteristic of the image. Various values of M is tested and the result is shown in Figure 5. As shown in the figure, better performance is achieved when more patches are extracted since more information is preserved. However, more patches will result in higher computational cost. In our experiment, we set M to 50,000 for the balance between performance and computational efficiency.

4.3. Evaluation on LIVE

Typically, NR-IQA methods are evaluated on the LIVE dataset for distortion-specific and non-distortion-specific experiments. For the former, only distorted images of the specific distortion are trained and tested, while for the latter

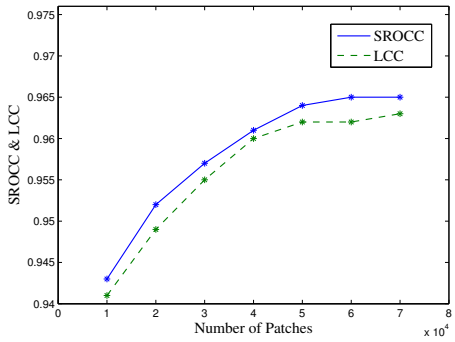


Figure 5. SROCC and LCC with respect to the number of Raw Patches for Local Feature Extraction

all images of all distortions are trained and tested together.

Table 1 shows our experimental results. The results of previous state-of-the-art NR-IQA and FR-IQA methods are also listed. Results for NR-IQA methods DIIVINE [16], BLIINDS-II [19], BRISQUE [14], CORNIA [28] and CNN [10] are taken from the original paper. CNN takes 60% of the dataset for training, 20% for validation and 20% for testing. All the others takes 80% of of the dataset for training, 20% for testing. Results for FR-IQA methods PSNR, SSIM and FSIM [29] are taken from [10]. The range of quality scores generated by FR-IQA algorithms are different from reference DMOS scores. In FR-IQA algorithms, the predicted scores are mapped to the certain range by a non-linear logistic function as follows:

$$Q_p = \beta_1 (\frac{1}{2} - \frac{1}{\exp(\beta_2 (Q - \beta_3))}) + \beta_4 Q + \beta_5 \quad (6)$$

Q is the quality score obtained by IQA algorithm and Q_p is the human perceptual quality (DMOS).

As shown in Table 1, the proposed SOM method achieves promising results in both distortion-specific (DS) experiments and non-distortion-specific (NDS) experiments. In DS experiments, our approach obtains promising results on all the five distortions, especially on white noise (WN), Gaussian blur (BLUR) and fast fading (FF). In NDS experiments, our approach achieves state-of-the-art result when compared to other NR-IQA and FR-IQA methods.

We owe the good performance of SOM to the combination of local characteristic feature and the semantic obviousness. The former has been adopted by many IQA algorithms and is able to represent image quality on pixel or signal level. The proposed semantic obviousness, however, measures image quality from human perspective rather than signal level.

We can notice from Table 1 that SOM performs relatively poor on JP2K and JPEG. Both JP2K and JPEG cause blockiness in local regions which leads to worse detection perfor-

SROCC	JP2K	JPEG	WN	BLUR	FF	ALL
PSNR	0.870	0.885	0.942	0.763	0.874	0.866
SSIM	0.939	0.946	0.964	0.907	0.941	0.913
PSIM	0.970	0.981	0.967	0.972	0.949	0.964
DIIVINE	0.913	0.910	0.984	0.921	0.863	0.916
BLIIDNS-II	0.929	0.942	0.969	0.923	0.889	0.931
BRISQUE	0.914	0.965	0.979	0.951	0.877	0.940
CORNIA	0.943	0.955	0.976	0.969	0.906	0.942
CNN	**0.952**	**0.977**	0.978	0.962	0.908	0.956
SOM	0.947	0.952	**0.984**	**0.976**	**0.937**	**0.964**
SOM-L	0.945	0.949	0.979	0.971	0.928	0.961
LCC	JP2K	JPEG	WN	BLUR	FF	ALL
PSNR	0.873	0.876	0.926	0.779	0.870	0.856
SSIM	0.921	0.955	0.982	0.893	0.939	0.906
PSIM	0.910	0.985	0.976	0.978	0.912	0.960
DIIVINE	0.922	0.921	0.988	0.923	0.888	0.917
BLIINDS-II	0.935	0.968	0.980	0.938	0.896	0.930
BRISQUE	0.923	0.973	0.985	0.951	0.903	0.942
CORNIA	0.951	0.965	0.987	0.968	0.917	0.935
CNN	**0.953**	**0.981**	0.984	0.953	0.933	0.953
SOM	0.952	0.961	**0.991**	**0.974**	**0.954**	**0.962**
SOM-L	0.950	0.957	0.988	0.969	0.945	0.958

Table 1. SROCC and LCC on LIVE. Italicized are FR-IQA

mance of our gradient-based detector. We'll try some other detectors in the future work.

Effectiveness of Semantic Obviousness

In Table 1, we denote the result with only local features as SOM-L. As we can see, if only local features are used, we'll get an obvious decrease in performance.

We design some extra experiments to further demonstrate the effectiveness of the the proposed semantic obviousness. In Figure 6(a) we show several distorted images from two different distortion types in LIVE dataset. The four images in the upper row are from the JPEG distortion and the four images in the bottom row are from fast fading (FF) distortion. In both rows, images are arranged with their quality scores decrease from (a) to (d). The top object detection scores of JPEG and FF distorted images are shown in Figure 6(b) and Figure 6(c) respectively. Only top-30 detection scores of each image are plotted for the convenience of our observation. As shown in the figure, images with a high quality score typically get higher object detection scores. We evaluate this correlation on the dataset, and find it's consistent for most of the cases. That's to say, the distribution of object scores contains useful information for image quality estimation. Based on such observation, semantic obviousness feature is constructed by the top object detection scores of the image.

4.4. Cross Dataset Evaluation

We test the generalization ability of our approach following previous NR-IQA methods [14, 10, 28]. We train our model on LIVE dataset and test the performance on TID2008. Only images of the four distortions shared by LIVE and TID2008 are tested in this experiment [10, 28].

	BRISQUE	CORNIA	CNN	**SOM**
SROCC	0.882	0.892	0.920	**0.923**
LCC	0.892	0.880	**0.903**	0.899

Table 2. The results of different methods on TID2008. The models are trained on LIVE dataset

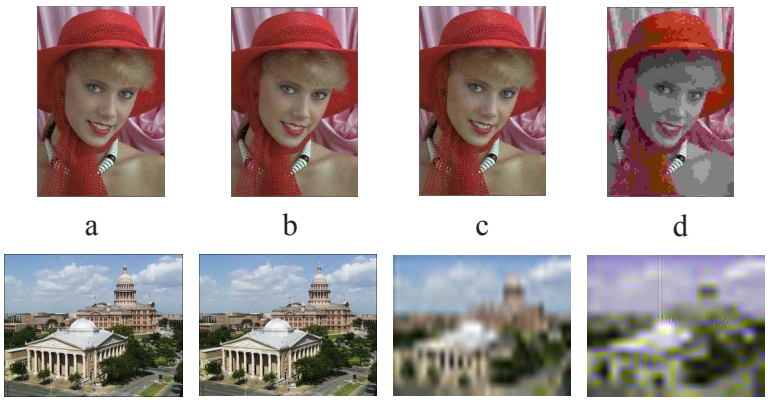
The DMOS scores of LIVE are in the range[1] $[0, 100]$, while MOS scores in TID2008 range from 0 to 9. We map the predicted scores to the MOS range by adopting the same method in [10, 28]. In each training-testing iteration, 80% of the data is used to fit the regression function (see Equation 6) and 20% is used for evaluation. The result obtained in 100 iterations is shown in Table 2. The proposed SOM achieves promising results on TID2008 with the model trained on LIVE dataset, which shows our method is generalizable.

4.5. Incorporation with Other IQA methods

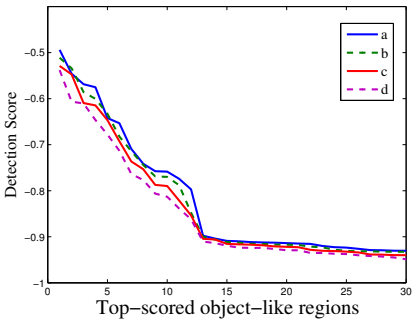
Some techniques we adopted in our framework can be easily generalized to other FR-IQA or NR-IQA algorithms. To demonstrate this, we combine the semantic obviousness metric with previous IQA methods, such as PSNR, SSIM [24] and BRISQUE [15]. For the full-reference algorithms PSNR and SSIM, we extract N top-scored object-like regions of each reference image. For each region, the corresponding region in the distorted image is extracted to obtain a predicted quality score. The scores of N regions is averaged to represent the image quality of the whole distorted image. The same non-linear logistic regression function (Equation 6) is used to map predicted scores to the range of DMOS. For the non-reference algorithm BRISQUE, we extract top N object-like regions for each distorted image and average the predicted scores of these regions to obtain the quality score of the whole distorted image. All parameters of the above algorithms are set to default. The results of the original algorithms and their object based versions are shown in Table 3. The experiments are performed on LIVE dataset with all the distorted images. The parameter N is set to 10, and better results may be obtained with a different value.

As shown in Table 3, by incorporation with generic object detection, the performance of existing methods can get obvious improvement. It's possible to obtain better boost of performance if we integrate object detection into these algorithms rather than simply averaging the scores of these detected regions.

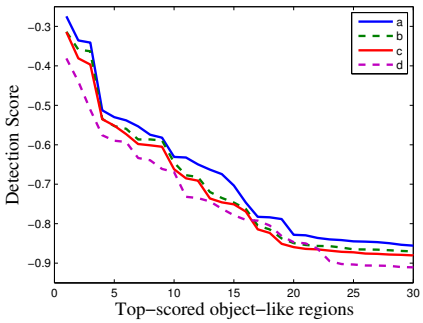
[1] Actually, in the latest release of LIVE dataset, DMOS scores are realigned and thus 21 scores exceed 100.



(a) Distorted images of JPEG and fast fading



(b) Detection scores of JPEG distorted images



(c) Detection scores of FF distorted images

Figure 6. Example of distorted images and object scores of their object-like regions

	SROCC	LCC
PSNR	0.872	0.867
PSNR+SOM	0.885	0.874
SSIM [24]	0.913	0.906
SSIM+SOM	0.933	0.922
BRISQUE [15]	0.940	0.942
BRISQUE+SOM	0.952	0.949

Table 3. The performance of three existing IQA algorithms without and with semantic obviousness metric

5. Conclusion

In this paper, we present a simple, effective and generalizable framework for general-purpose non-reference image quality assessment. Our semantic obviousness based algorithm measures the image quality from both signal level and semantic sense. The predicted image quality of our method demonstrates high consistency with human perceptual quality. Our method outperforms the state-of-the-art NR-IQA methods and is comparable to the FR-IQA methods. We also demonstrate that the techniques adopted in our approach can be incorporated to boost the performance of existing FR-IQA and NR-IQA methods. More powerful object detectors can be exploited to improve the performance.

6. Acknowledgement

This work was supported in part to Prof. Houqiang Li by 973 Program under contract No. 2015CB351803, NSFC under contract No. 61325009 and No. 61390514, and in part to Dr. Wengang Zhou by NSFC under contract No. 61472378 and the Fundamental Research Funds for the Central Universities under contract No. WK2100060014 and WK2100060011.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, pages 73–80. IEEE, 2010.

[2] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *TPAMI*, 34(11):2189–2202, 2012.

[3] D. M. Chandler and S. S. Hemami. Vsnr: A wavelet-based visual signal-to-noise ratio for natural images. *TIP*, 16(9):2284–2298, 2007.

[4] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *CVPR*, pages 3286–3293. IEEE, 2014.

[5] A. Coates, A. Y. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *ICAIS*, pages 215–223, 2011.

[6] M. P. Eckert and A. P. Bradley. Perceptual quality metrics applied to still image compression. *Signal Processing*, 70(3):177 – 200, 1998.

[7] I. Endres and D. Hoiem. Category-independent object proposals with diverse ranking. *TPAMI*, 36(2):222–234, Feb 2014.

[8] L. He, D. Tao, X. Li, and X. Gao. Sparse representation for blind image quality assessment. In *CVPR*, pages 1146–1153. IEEE, 2012.

[9] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113. IEEE, 2009.

[10] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *CVPR*, pages 1733 – 1740. IEEE, 2014.

[11] D.-O. Kim, H.-S. Han, and R.-H. Park. Gradient information-based image quality metric. *IEEE Transactions on Consumer Electronics*, 56(2):930–936, 2010.

[12] E. C. Larson and D. M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, 2010.

[13] Q. Ma and L. Zhang. Saliency-based image quality assessment criterion. In *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, pages 1124–1133. Springer, 2008.

[14] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *TIP*, 21(12):4695–4708, 2012.

[15] A. K. Moorthy and A. C. Bovik. A two-step framework for constructing blind image quality indices. *Signal Processing Letters*, 17(5):513–516, 2010.

[16] A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *TIP*, 20(12):3350–3364, 2011.

[17] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barbba. Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric. In *ICIP*, volume 2, pages II169–II172. IEEE, 2007.

[18] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009.

[19] M. A. Saad, A. C. Bovik, and C. Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *TIP*, 21(8):3339–3352, 2012.

[20] H. R. Sheikh, A. C. Bovik, and G. De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *TIP*, 14(12):2117–2128, 2005.

[21] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. Live image quality assessment database release 2, 2005.

[22] H. Tang, N. Joshi, and A. Kapoor. Learning a blind measure of perceptual image quality. In *CVPR*, pages 305–312. IEEE, 2011.

[23] Y. Tong, H. Konik, F. Cheikh, and A. Tremeau. Full reference image quality assessment based on saliency map analysis. *JIST*, 54(3):30503–1, 2010.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004.

[25] Z. Wang, L. Lu, and A. C. Bovik. Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication*, 19(2):121 – 132, 2004.

[26] W. Xue, L. Zhang, X. Mou, and A. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *TIP*, 23(2):684–695, Feb 2014.

[27] P. Ye and D. Doermann. No-reference image quality assessment based on visual codebook. In *ICIP*, pages 3089–3092. IEEE, 2011.

[28] P. Ye, J. Kumar, L. Kang, and D. Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *CVPR*, pages 1098–1105. IEEE, 2012.

[29] L. Zhang, D. Zhang, and X. Mou. Fsim: a feature similarity index for image quality assessment. *TIP*, 20(8):2378–2386, 2011.