Real-time No-Reference Image Quality Assessment based on Filter Learning

Peng Ye, Jayant Kumar, Le Kang, David Doermann
Institute for Advanced Computer Studies
University of Maryland, College Park, MD, USA
{pengye,jayant,lekang,doermann}@umiacs.umd.edu

Abstract

This paper addresses the problem of general-purpose No-Reference Image Quality Assessment (NR-IQA) with the goal of developing a real-time, cross-domain model that can predict the quality of distorted images without prior knowledge of non-distorted reference images and types of distortions present in these images. The contributions of our work are two-fold: first, the proposed method is highly efficient. NR-IQA measures are often used in real-time imaging or communication systems, therefore it is important to have a fast NR-IQA algorithm that can be used in these real-time applications. Second, the proposed method has the potential to be used in multiple image domains. Previous work on NR-IQA focus primarily on predicting quality of natural scene image with respect to human perception, yet, in other image domains, the final receiver of a digital image may not be a human.

The proposed method consists of the following components: (1) a local feature extractor; (2) a global feature extractor and (3) a regression model. While previous approaches usually treat local feature extraction and regression model training independently, we propose a supervised method based on back-projection, which links the two steps by learning a compact set of filters which can be applied to local image patches to obtain discriminative local features. Using a small set of filters, the proposed method is extremely fast. We have tested this method on various natural scene and document image datasets and obtained state-of-the-art results.

1. Introduction

With the advancement of digital imaging, there has been a tremendous growth in using digital images for representing and communicating information. In such an environment, it is critical to have good image quality assessment methods to help maintain, control and enhance the quality of the digital images.

The goal of objective image quality assessment (IQA) is to build a computational model that can accurately predict the quality of digital images with respect to human perception or other measures of interest. Based on the availability of reference images, objective IQA approaches can be classified into: full-reference (FR), no-reference (NR) and reduced-reference (RR) approaches.

This paper addresses the most challenging category of objective IQA methods – NR-IQA, which evaluates the quality of digital images without access to reference images [3, 11, 12, 13, 16, 19]. More specifically, we develop a general-purpose NR-IQA algorithm which does not require prior knowledge of the types of distortions.

NR-IQA has long been considered as one of the most difficult problems in image analysis [20]. Without knowledge of the reference image and the type of distortion, this problem may seem difficult, but recently, significant progress has been made in the field. State-of-the-art general-purpose NR-IQA systems [12, 13, 16, 22] have been shown to outperform FR measures Peak Signal-to-Noise ratio (PSNR) and Structural Similarity Index Measure (SSIM) on standard IQA dataset.

1.1. Motivation

Speed is an important issue for NR-IQA systems since NR-IQA measures are often used in real-time imaging or communication systems. Algorithms that rely on computationally expensive image transforms [13, 16] often can not be used in these applications. By extracting image quality features directly in spatial domain, recent algorithms CORNIA [22] and BRISQUE[12] have greatly accelerated this process while maintaining high prediction accuracy. By using a compact set of filters, our method can further accelerate the process.

Previous works on NR-IQA have focused primarily on natural scene image and image quality is defined with respect to human perception. Very limited work has been done for NR-IQA for other types of images, such as camera-captured or scanned document images. Document IQA has been found to be very useful in many document image processing applications. For example, depending on the level

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.132

of degradation, the performance of modern OCR software may suffer. Document IQA may help to automatically filter pages with low predicted OCR accuracy or guide the selection of document image enhancement methods.

Conventional image quality measures developed for natural scene images may not work well for document images since document images have very different characteristics than natural scene images. For example, most document images are gray-scale or binary consisting of black text and white background. Building a NR-IQA system that can be adapted to images with different characteristics is a challenging problem.

To sum up, the objective of this work is: first, to develop a fast NR-IQA method that can be used in real-time systems and second, to develop a general learning-based framework that can be applied to various different image domains.

1.2. Related work

Natural Scene Statistics for NR-IQA

Natural Scene Statistics (NSS) based approaches have been successfully applied to IQA for natural scene images. These methods are based on the following observations: first, when images are properly normalized or transferred to some transform domains (e.g. DCT or wavelet domain), local descriptors (e.g. normalized intensity values, wavelet coefficients, etc), can be modeled by some parametric distributions; second, the shape of these distributions are very different for non-distorted and distorted images. These fundamental observations form basis of many recent IQA approaches [12, 13, 16]. These methods differ from each other primarily in how the local descriptors are extracted. For example, in DIIVINE [13] local descriptors are extracted in wavelet domain. Cosine transform coefficients based descriptors are used in BLIINDS-II [16]. BRISQUE [12] directly models the normalized image pixel value using generalized Gaussian distributions (GGD) and models product of neighboring pixels by asymmetric generalized Gaussian distributions (AGGD). The success of these methods rely largely on how local features are computed, therefore handcraft features designed specifically for a particular domain are often used. This limits the application of these methods in other image domains.

Feature Learning

Instead of using hand-craft local descriptors, the proposed approach is based on feature learning. The goal is to learn local features whose distributions possess discriminative shapes for distorted and non-distorted images. Unsupervised feature learning has been explored in CORNIA [22], where the local descriptors are encoded using codeword that are learned in an unsupervised way. The success of this method relies on using a large set of codeword (usually in order of thousands), which can capture different aspects of distortions. As was shown in [22], when only a small set of codewords are used, the performance of this method drops significantly. Our method can be considered as a supervised extension of CORNIA, where instead of using a large redundant set of filters, the proposed method learns a compact set of filters in a supervised way. Using a small set a filters, our feature extraction process is much faster and more memory efficient.

The proposed supervised filter learning method is closely related to supervised dictionary learning for image classification. Earlier methods for dictionary learning focused on reconstruction of signals and ignored label information. To learn a more compact and discriminative dictionary, learning approaches that jointly optimize both a reconstructive and a discriminative criterion have been developed [10, 21, 7]. Unlike conventional supervised dictionary learning, which requires that the linear combination of the learned atoms in dictionary should be able to well represent image patches, we do not have this constraint in our supervised filter learning process. In fact, it will be shown later that the functionality of filter for NR-IQA and codeword for image classification are very different.

Supervised filter learning has also been explored by Jain and Karu in [6] for texture classification, where feature extraction and classification tasks are performed by a neural network. The learned filters are weight vectors in the first layer of the network. Our work is along the same lines of learning a compact set of filters using a back-propagation approach but differs in final stage where we perform support vector regression (SVR) using learned filters for predicting image quality.

1.3. Our approach

A typical NR-IQA system may consist of the following three components (1) a local feature extractor; (2) a global feature extractor, which summarizes the distribution of local features and (3) a regression model. Previous approaches usually treat local feature extraction and regression model training independently. We propose a supervised method based on back-projection, which links these two steps by jointly optimizing the prediction model and the local feature extractor. The learned compact set of filters when applied to local image patches yields more discriminative features. Additionally, due to a significant reduction in the number of filters, the proposed method achieves much better time performance in comparison to previous approaches. An overview of our system is shown in Fig. 1.

2. Feature extraction

In this section, we discuss how a set of linear filters are used to obtain global features. Suppose an image is represented by a set of local descriptors, where these local descriptors are normalized raw image patches:

$$X = [x_1, x_2, ..., x_N] \in R^{d \times N}$$

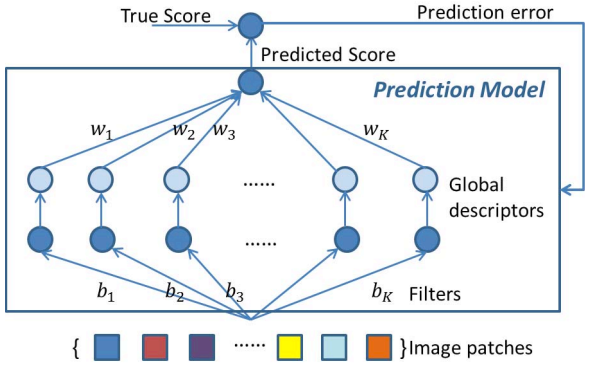


Figure 1: Overview of the proposed approach with linear SVR.

where the column vector x_i denotes the i-th local descriptor of the image. The normalization is performed by subtracting the local mean value from each patch, and dividing it by its standard deviation.

A set of filters are represented by $B = [b_1, ..., b_K] \in R^{d \times K}$, where the column vector b_i ($||b_i||_{l2}^2 = 1$) denotes the i-th filter and K is the number of filters.

2.1. Local feature encoding

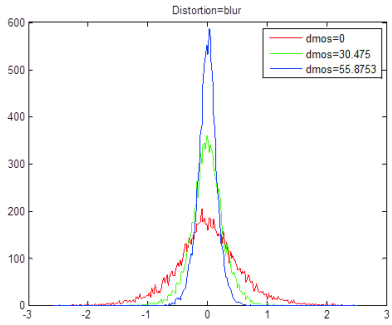
The first step in this work-flow is local feature encoding using linear filters. Specifically, each local descriptor is encoded by its responses to the set of linear filters (i.e. inner product between local descriptors and filters). We have an image level representation matrix as follows:

$$\Omega = B^T \times X = \begin{pmatrix} b_1 \cdot x_1 & b_1 \cdot x_2 & \cdots & b_1 \cdot x_N \\ b_2 \cdot x_1 & b_2 \cdot x_2 & \cdots & b_2 \cdot x_N \\ \vdots & \vdots & \ddots & \vdots \\ b_K \cdot x_1 & b_K \cdot x_2 & \cdots & b_K \cdot x_N \end{pmatrix} \quad (1)$$

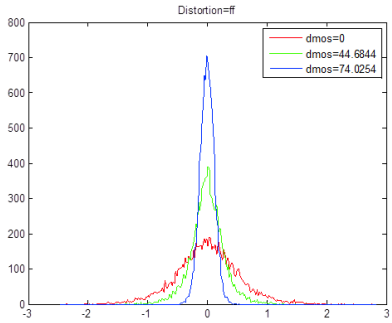
Examples of distributions of filter responses from images with different types and levels of distortions are shown in Fig. 2. We can see from this figure that with properly learned filters, statistics extracted from these distributions can be good indicators of image quality. The *filter* here is similar to the *codeword* in the image classification literature in that both are used for local feature encoding. However, their functionalities are very different. For example, in the object recognition problem, whereas a *codeword* resembles a part of the object, and the maximal response to each codeword indicates its presence or absence in the image. In our problem, both maximal and minimal responses are informative and important for prediction task. More generally, we are interested in characterizing the entire distribution of the filter responses.

2.2. Summarizing statistics

Statistics which summarize the distribution of local features are extracted as global descriptors. Specifically, we



(a) Gaussian Blurring



(b) Fast Fading

Figure 2: Examples of filter respones for different types and levels of distortions. (High DMOS indicates low quality)

use the maximal and minimal value of filter responses for describing the effect of filters on the image. The image level descriptor of X can be written as:

$$Z = [max(\Omega)^T, min(\Omega)^T] \quad (2)$$

where max and min are operated on each row of Ω and superscript T means transpose. Other statistics which summarize the distribution of filter responses can also be explored, such as skewness and kurtosis. One can also use a parametric model, for example, BRISQUE [12] models normalized intensity value using GGD where the shape and scale parameters are used as features. Although the minimal and maximal value of filter responses may not be accurate in characterizing the shape of distribution, as is shown in Fig. 2, for discriminating images with high and low quality, they are fairly good indicators, in addition to being efficient to compute.

Combining the two steps above, we have $Z = \phi(X, B)$ where $Z \in R^{2K \times 1}$ with the first K elements corresponding to maximal responses and the last K elements corresponding to minimal responses.

3. Supervised filter learning

In the previous section, we have introduced the use of a set of filters for obtaining global descriptors of an image. In

this section, we discuss how these filters are learned so that the corresponding global descriptors are good for NR-IQA tasks.

3.1. Problem formulation

Suppose we have n training images and the k-th training image is denoted as X_k with corresponding feature vector denoted as Z_k. Its regression target, i.e, the true quality score, is denoted as y_k. In this paper, we use linear ϵ-Support Vector Machine Regression (ϵ-SVR) for training. The prediction function takes the form: $f(Z_k, w) = \sum_{i=1}^{2K} w_i Z_k(i) + w_0$, where $Z_k(i)$ is the i-th element in Z_k and w is learned by minimizing the sum of a loss function and a regularization term:

$$min_w\{\sum_{k=1}^{n} L(y_k, f(Z_k, w)) + \lambda_1 ||w||_{l_2}^2\}$$

where L is the ϵ-insensitive loss function described by:

$$L(y, \hat{y}) = \begin{cases} 0 & \text{if } |y - \hat{y}| \leq \epsilon \\ |y - \hat{y}| - \epsilon & \text{otherwise} \end{cases} \quad (3)$$

In the above formulation, the prediction model is trained with the set of filters fixed. Our supervised filter learning method jointly optimizes the prediction model and the set of filters. The objective function of this optimization problem is defined as follows:

$$\begin{gathered} C(B, w, \{X_k\}_{k=1}^n) = \sum_{k=1}^n L(y_k, f(\phi(X_k, B), w)) \\ +\lambda_1 ||w||_{l_2}^2 + \lambda_2 avecorr(B) \\ subject\ to\ ||b_i|| = 1, i = 1, ..., K \end{gathered} \quad (4)$$

where λ_1 is a balancing factor of the regularization term in the prediction model and $avecorr(B) = \frac{1}{K-1}\sum_{i=1}^{K}\sum_{j:j\neq i} < b_i, b_j >$ is the average correlation of one filter with every other filters. This correlation penalty term is added to avoid learning highly correlated filters. Furthermore, in order to avoid the over-fitting problem and regularize the search space of the optimal filters, we add the constraint that $||b_i|| = 1 (i = 1, ..., K)$.

Optimal B and w is given by

$$(B^*, w^*) = argmin_{B,w} C(B, w, \{X_k\}_{k=1}^n)$$

This optimization problem can be solved by optimizing alternatively over B and w. The initial set of filters are obtained by performing k-means clustering on a set of local features. When B is fixed, optimal w can be found using a standard SVR program [4]. Given w fixed, we can apply stochastic gradient descent (SGD) to find the optimal B.

3.2. Optimizing B

Computing Gradient

The SGD process requires us to compute the gradient of C with respect to B. The gradient is computed as follows.

$$\frac{\partial C}{\partial B} = \sum_{k=1}^{n} \frac{\partial L}{\partial f_k}\frac{\partial f_k}{\partial Z_k}\frac{\partial Z_k}{\partial B} + \lambda_2 \frac{\partial avecorr(B)}{\partial B} \quad (5)$$

where $f_k = f(Z_k, w)$. When linear SVR is used, we can compute the derivative of the objective function as follows:

$$\frac{\partial C}{\partial b_i} = \sum_{k=1}^{n} \frac{\partial L}{\partial f_k}(w_i x_{max,i}^k + w_{i+K} x_{min,i}^k) + \lambda_2 \frac{1}{K-1}\sum_{j:j\neq i} b_j \quad (6)$$

where $x_{max,i}^k = argmax_{x_l \in X_k}(b_i \cdot x_l)$ and $x_{min,i}^k = argmin_{x_l \in X_k}(b_i \cdot x_l)$. Superscript k is the index of the training image and $x_l \in X_k$ means x_l is a local feature vector from image X_k. $\frac{\partial L}{\partial f_k}$ accounts for the prediction error of the k-th training image. The loss function defined in Eq. 3 is not differentiable, so we use the Huber loss as an approximation for computing the gradient. Details of this computation process is provided in the supplementary material.

Stochastic Gradient Descent

Our optimization problem has the constraint that $||b_i||_{l_2}^2 = 1$, so we perform SGD on the unit sphere. This can be done by projecting the gradient on the tangent plane of the sphere. We describe the SGD process for optimizing a given filter as follows:

1. Permute the training images randomly, set $k = 1$ and initialize b^1.

2. Compute the gradient $g_k = \nabla_b C^k(b)|_{b=b^k}$, where $C^k = L(y_k, f(Z_k, w)) + \lambda_2 avecorr(b^k)$ and b^k is the value of the filter at the k-th iteration.

3. Project g_k on the tangent plane of the unit sphere at b^k, $h_k = g_k - (g_k \cdot b^k)b^k$ and normalize it, $n_k = h_k/|h_k|$.

4. Update b^k with $b^{k+1} = b^k cos(r_k) + n_k sin(r_k)$, where r_k is the learning rate at the k-th iteration. Set $k = k + 1$.

5. Go back to step 2 and repeat the process until the maximal number of iterations is reached.

We optimize a set of filters B by iterating the above process. At each step, we simultaneously update all filters in B and update the objective function with the new set of filters.

Early Stopping

The back projection based method may suffer from the over-fitting problem. In order to avoid over-fitting, we adopt the early stopping criteria, proposed in [15]. Specifically, we divide the training data into a training set and a validation set. In each iteration of the optimization process, we train on the training set and test on the validation set. The training process is terminated as soon as the error on validation set satisfies our early stopping rule. The set of filters which gives the best performance on the validation set is chosen as the output of the optimization process. In NR-IQA problems, the linear correlation coefficient (LCC) is usually used as an evaluation measure. So we define the following generalization loss based on the

	LIVE	TID08	SOC	Newspaper
patch size	7	5	7	5
ϵ	1	0.5	0.1	0.2
$\lambda 2$	1	–	1	1
r_0	0.001	–	0.002	0.005

Table 1: Parameters used in our experiments.

LCC. The generalization loss at the k-th iteration is give by $GL(k) = 100((1 - corr(k))/(1 - corr_{opt}(k)) - 1)$, where $corr(k)$ is the LCC on validation set at the k-th iteration and $corr_{opt}(k) = max_{k' \leq k} corr(k')$. Two stopping rules are used:

(**Rule 1**) $GL(k) > \alpha$ for some $\alpha > 0$;

(**Rule 2**) *corr* decreases in consecutive l iterations.

Training terminates if Rule 1 or Rule 2 is true.

4. Experiment

4.1. Protocol

We refer to the supervised filter learning based method with k filters as *SFk* and the corresponding unsupervised method as *CBk*. For example *CB100* refers to the method using 100 filters obtained from kmeans clustering. *SFk-BS7* (*CBk-BS7*), *SFk-BS5* (*CBk-BS5*) refer to methods with patch size 5×5 and 7×7 respectively.

For different experiments, different experimental settings may be used. We first introduce the experimental settings that are fixed for all experiments: (1) We use linear ϵ-SVR for regression with λ_1 fixed to one. (ϵ may change for different experiments). (2) Around 8000-10000 patches are extracted from each image. (3) In cross-validation experiments, 80% of the data is used for training and the remaining 20% is used for testing. (4) For the *SF* method, around 1/4 of the training data is used as a validation set. (5) The minimal number of iterations of the optimization process is set to 50 and the maximal number of iterations is set to 250. After 50 iterations, if one of the two early stopping rules is true or the maximal number of iterations is reached, we terminate the filter training process. (6) Parameters in the early stopping rules are $\alpha = 10, l = 10$. (7) The learning rate in SGD is $r_t = \frac{r_0}{\sqrt{1+t/N}}$, where t is the number of iteration, N is the number of training samples and r_0 is the initial learning rate. (8) For evaluating IQA measures, we follow common practice and use linear correlation coefficient (LCC) and Spearman rank order correlation (SROCC) as our evaluation measures.

There may be parameters which are set differently for different experiments. Table 1 summarizes the choice of parameters in different experiments.

4.2. Experiments on Natural Scene Image

4.2.1 Dataset

The following two IQA datasets were used to test the proposed method on natural scene images.

(1) *LIVE IQA dataset*: LIVE IQA dataset [17, 18] consists of images with five types of distortions - JPEG2k, JPEG, white Gaussian noise (WN), Gaussian blurring (BLUR) and fast fading channel distortion (FF) derived from 29 non-distorted images. The *Differential Mean Opinion Score* (DMOS) associated with distorted images is provided. DMOS is generally in the range $[0, 100]$, where lower DMOS indicates higher quality. (2) *TID2008 dataset*: 25 reference images and 1700 distorted images derived from them with 17 different distortions at 4 levels are included in TID2008 [14]. In this paper, we consider four of the 17 distortions which are most likely to occur in image processing systems, including Additive Gaussian noise (WN), JPEG compression (JPEG), JPEG2000 compression (JPEG2K) and Gaussian blur (BLUR). Each distorted image is associated with a Mean Opinion Score (MOS). Higher values of MOS (0 - minimal, 9 - maximal) correspond to higher visual quality of the image.

4.2.2 Evaluation

Test on LIVE

In the first set of experiments, we randomly split the LIVE dataset into training and testing parts. At each iteration we train the codebook and regression model on the training set and evaluate the trained model on the testing set. This process was repeated for each distortion and also for the complete set of all distortions. 100 repeated experiments were performed and the resulting median value of LCC and SROCC are shown in Table 2 and Table 3 respectively[1]. Results can also be found for two FR measures – PSNR and SSIM and two other NR measures – CORNIA [22] and BRISQUE [12] in the tables. In [12], BRISQUE was trained using SVR with RBF kernel. Since linear SVR is used in our method and CORNIA, in our experiment, we compare with BRISQUE-L, which is trained with linear SVR using coarsely optimized parameters.

It can be seen that *SF100* outperforms PSNR and is comparable to the SSIM. CORNIA is equivalent to *CB10000*, which is based on a large set of filters. Using only 100 filters, the performance of *SF100* is only slightly lower than CORNIA and BRISQUE. As will be shown later in this section, with a little loss in prediction accuracy, we can achieve significant improvement in speed.

The optimization process on the entire LIVE from the first 51 iterations averaged over all the 100-fold experiments

[1] Only distorted images are used in testing.

	JP2K	JPEG	WN	BLUR	FF	ALL
PSNR	0.872	0.885	0.941	0.764	0.875	0.867
SSIM	0.939	0.946	0.965	0.909	0.941	0.914
CORNIA	0.943	0.955	0.976	0.969	0.906	0.942
BRISQUE-L	0.910	0.961	0.987	0.947	0.903	0.929
CB100	0.915	0.846	0.953	0.946	0.878	0.839
SF100	0.924	0.928	0.962	0.961	0.879	0.920

Table 2: Median SROCC with 100 iterations of experiments on the LIVE dataset. (*Italicized* algorithms are NR-IQA algorithms, others are FR-IQA algorithms.)

	JP2K	JPEG	WN	BLUR	FF	ALL
PSNR	0.873	0.874	0.928	0.774	0.869	0.855
SSIM	0.920	0.955	0.982	0.891	0.939	0.906
CORNIA	0.951	0.965	0.987	0.968	0.917	0.935
BRISQUE-L	0.910	0.966	0.992	0.943	0.929	0.929
CB100	0.918	0.843	0.970	0.947	0.878	0.821
SF100	0.929	0.940	0.978	0.960	0.888	0.921

Table 3: Median LCC with 100 iterations of experiments on the LIVE dataset. (*Italicized* algorithms are NR-IQA algorithms, others are FR-IQA algorithms.)

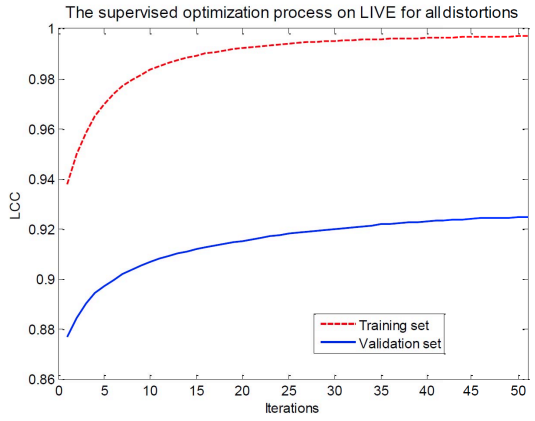


Figure 3: Optimization process of the first 51 iterations on training set and validation set (average LCC from 100 fold experiments on LIVE).

is shown in Fig. 3. It can be seen that the training process converges nicely.

We also compare the supervised filter learning method with unsupervised filter learning *CB100-BS7*. When only 100 filters are used, the prediction performance can be significantly boosted by adopting the supervised approach.

Supervised filter training on LIVE, test on TID2008

To show that the proposed method does not depend on any particular dataset, we train filters on the LIVE dataset and test it on the TID2008 dataset. Using the fixed set of filters trained on LIVE, in each iteration of the experiment, we split the TID2008 dataset into non-overlapping training and testing part and perform training and testing on each

	JP2K	JPEG	WN	BLUR	ALL
PSNR	0.884	0.922	0.920	0.928	0.832
SSIM	0.961	0.934	0.854	0.958	0.908
CORNIA	0.932	0.926	0.907	0.908	0.884
CB100	0.919	0.881	0.865	0.869	0.754
CB200	0.926	0.907	0.864	0.882	0.800
SF100	0.917	0.913	0.899	0.911	0.830
SF200	0.928	0.944	0.897	0.902	0.873

Table 4: Median SROCC with 1000 iterations of experiments on the TID2008 dataset.

	JP2K	JPEG	WN	BLUR	ALL
PSNR	0.907	0.928	0.944	0.912	0.789
SSIM	0.971	0.963	0.827	0.953	0.907
CORNIA	0.939	0.960	0.898	0.906	0.924
CB100	0.932	0.905	0.859	0.855	0.797
CB200	0.934	0.933	0.854	0.882	0.850
SF100	0.932	0.939	0.894	0.910	0.868
SF200	0.937	0.971	0.891	0.899	0.901

Table 5: Median LCC with 1000 iterations of experiments on the TID2008 dataset.

	SF100-BS7	*SF100-BS5*	BRISQUE	CORNIA
Time	0.037	0.058	0.119	0.332

Table 6: Feature extraction time (in seconds).

part. This experiment was repeated 1000 times and the median values of SROCC and LCC are reported in Table 4 and Table 5. For a specific distortion category, the set of filters are trained on corresponding subset in LIVE. Comparing results obtained using *CB100-BS5*/*CB200-BS5* and *SF100-BS5*/*SF200-BS5*, we can see great performance improvement by using a supervised feature learning strategy. The performance of *SF200-BS5* is comparable to CORNIA and outperforms the FR measure – PSNR.

Speed Test

Since IQA measures are often used in real-time imaging or communication systems, speed is an important issue determining whether an IQA measure can be used in these applications. We test the speed of our method with 100 filters and two other recent fast NR-IQA measures. All three methods are implemented in Matlab and are tested on a SunFire x4170 with 2.80GH processor. We consider only feature extraction time and results are shown in Table 6. It is clear that *SF100* is much faster than other NR-IQA methods.

We further decompose the feature extraction process into two steps (1) extracting non-overlapping image patches and (2) extracting local and global feature based on Eq. 1 and 2. We then test their speeds respectively. For *SF100-BS7*, the time taken for these two steps are around 0.027 seconds and 0.01 seconds respectively. *SF100-BS5* is slower than *SF100-BS7* mainly because it takes longer time to extract distinct *5-by-5* patches compared to *7-by-7* patches.

4.3. Experiments on Document Image

Instead of predicting human perceived image quality, for document IQA (Doc-IQA), we are interested in predicting the OCR accuracy with respect to a specific OCR software, which has been found useful in many document image applications.

4.3.1 Dataset

To test the proposed method on document images, the following two document datasets are used.

(1) *SOC dataset*: *Sharpness-OCR-Correlation*(SOC) dataset contains camera-captured document with blur. 25 non-distorted images in this dataset are taken from two freely available datasets - *University of Washington Dataset* [5] and *Tobacco Database* [9]. For each document, multiple photos were taken from a fixed distance to capture the whole document, but the camera was focused at varying distance to generate a series of images with focal blur. A total of 25 such sets, each consisting of 6-8 high-resolution images (dimension: 3264×1840) were created using an Android phone with an 8 mega-pixel camera. ABBYY Fine Reader was used to obtain the OCR results and OCR accuracy were computed using ISRI-OCR evaluation tool [1]. OCR accuracy is in the range from 0 to 1. Details of this dataset is available in [8].

(2) *Newspaper dataset*: This dataset contains a total of 521 text zone images from a collection of gray scale newspaper images with machine-printed English and Greek text [2]. Each text zone contains more than 30 characters. OCR results obtained using ABBYY Fine Reader and associated OCR accuracy are available for each image. Examples of images from this dataset are shown in Fig. 4.

4.3.2 Evaluation

Test on SOC dataset

Experimental results on the SOC dataset are shown in Table 7[2]. This dataset contains mainly blur distortion. Blur is a relatively easy distortion and as is shown Table 2 and Table 3, all IQA measures performs fairly well on the blur category in LIVE. However, when tested on document image, we observe a significant drop in performance for all IQA measures under comparison. This is partly due to the fact that OCR accuracy may not be consistent with human perception. Consider two document images with the same level of blur distortion, they are expected to have similar OCR accuracy. However, it is hard to tell which image has higher OCR accuracy based on their visual appearances, since OCR accuracy is also related to the content of the document. This explains why we obtained LCC around 90%, while SROCC, which measures the monotonicity of the prediction model is much lower than the corresponding LCC. Only slight improvement has been observed using supervised filter based approach over unsupervised approach on the SOC dataset. This may be due to the fact that OCR accuracy, i.e. the regression target, for this dataset is highly imbalanced. Our method replies heavily on training, therefore the "quality" of training set is important.

[2]When the number of filters is very small, the initial set of filters are fairly unique. Therefore we set λ_2 as zero for *SF10* on the SOC dataset.

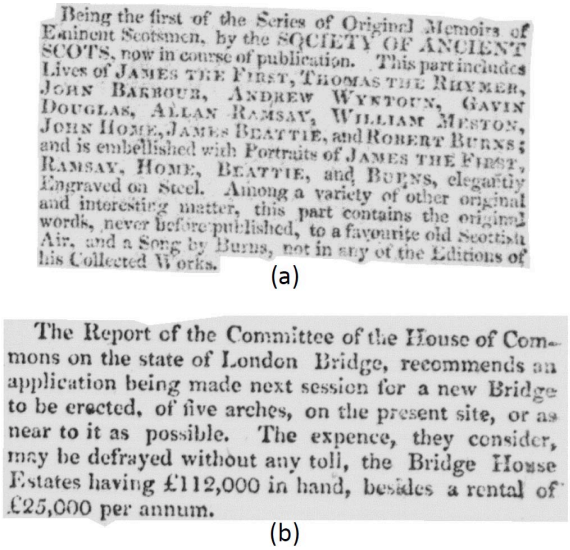


Figure 4: Examples of images from the newspaper dataset. Character level OCR accuracy (a) 0.097 (b) 0.958.

	BRISQUE-L	*CB10*	*CB100*	*CB10000*	*SF10*	*SF100*
LCC	0.904	0.873	0.927	0.937	0.886	0.927
SROCC	0.836	0.736	0.839	0.862	0.780	0.854

Table 7: Median LCC and SROCC with 100 iterations of experiments on the SOC dataset.

	BRISQUE-L	*CB200*	*CB5000*	*SF200*
LCC	0.722	0.692	0.751	0.735
SROCC	0.709	0.640	0.725	0.708

Table 8: Median LCC and SROCC with 100 iterations of experiments on the Newspaper dataset.

Test on Newspaper dataset

Experimental results on the Newspaper dataset are shown in Table 8. It can be seen that *SF200* significantly outperforms *CB200* and is comparable to BRISQUE-L. The local patch extraction on the newspaper dataset was performed only on text regions.

4.4. Discussions

It usually takes several runs to get optimal result using SGD. In our experiments, we only used single-run SGD, therefore, there is no guarantee that we have found global

optimal filters. To further improve the performance, we can repeat the iterative optimization process several times and choose the final learned model as the one which gives the best performance on the validation set.

When the initial set of filters obtained in an unsupervised manner has been able to well capture the distortion properties, supervised learning may not be very helpful. Using our approach, one can decide whether supervised extension is needed for different filter based on prediction tasks. Boosting over a large set of filters, say 10000, will not lead to much performance improvement. The development of a good Doc-IQA dataset is critical for developing any NR-IQA measures in document image domain. Available Doc-IQA datasets suffer from imbalance problem. It will be our future work to develop a high-quality Doc-IQA dataset to facilitate Doc-IQA research.

5. Conclusions

We have presented a supervised filter learning based algorithm for general-purpose NR-IQA. Compared to previous approaches, the proposed method has the following advantages (1) using a small set of learned filters and operating directly on raw image patches, this method is extremely fast; (2) unlike methods relying on hand-craft local descriptors which may not be generalizable for different image domains, our method is based on supervised feature learning, which has the potential to be adapted to different image domains. From our experiments, we conclude that a compact set of learned filters can achieve the same accuracy as by using a large number of unsupervised filters while reducing the the computation time significantly.

References

[1] ISRI-OCR evaluation tool UNLV/ISRI. *Available Online: http://code.google.com/p/isri-ocr-evaluation-tools/*, 2010.

[2] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher. Historical document layout analysis competition. In *ICDAR*, pages 1516–1520, Beijing, China, Sept. 2011.

[3] T. Brandão and M. P. Queluz. No-reference image quality assessment based on DCT domain statistics. *Signal Processing*, 88:822–833, Apr. 2008.

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[5] I. Guyon, R. M. Haralick, J. J. Hull, and I. T. Phillips. Data sets for OCR and document image understanding research. In *In Proceedings of the SPIE - Document Recognition IV*, pages 779–799. World Scientific, 1997.

[6] A. K. Jain and K. Karu. Learning texture discrimination masks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(2):195–205, Feb. 1996.

[7] Z. Jiang, Z. Lin, and L. S. Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1697–1704. IEEE, 2011.

[8] J. Kumar, P. Ye, and D. Doermann. DIQA: Document image quality assesment datasets. Available Online: http://lampsrv02.umiacs.umd.edu/projdb/project.php?id=73.

[9] D. Lewis. Building a test collection for complex document information processing. In *29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 665–666, 2006.

[10] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *NIPS*, pages 1033–1040, 2008.

[11] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi. Perceptual blur and ringing metrics: application to JPEG2000. *Signal Processing: Image Communication*, 19(2):163–172, 2004.

[12] A. Mittal, A. Moorthy, and A. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, PP(99):1, 2012.

[13] A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, Dec. 2011.

[14] 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 Radio Electronics*, 10:30–45, 2009.

[15] L. Prechelt. Early stopping - but when? In *Neural Networks: Tricks of the Trade, volume 1524 of LNCS, Chapter 2*, pages 55–69. Springer-Verlag, 1997.

[16] 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*, 21(8):3339 –3352, Aug. 2012.

[17] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006.

[18] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. LIVE image quality assessment database release 2. Online, http://live.ece.utexas.edu/research/quality.

[19] H. Tang, N. Joshi, and A. Kapoor. Learning a blind measure of perceptual image quality. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 305 – 312, 2011.

[20] Z. Wang, A. C. Bovik, and L. Lu. Why is image quality assessment so difficult? In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–3313 –IV–3316, May 2002.

[21] J. Yang, K. Yu, and T. Huang. Supervised translation-invariant sparse coding. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3517 –3524, 2010.

[22] P. Ye, J. Kumar, L. Kang, and D. Doermann. Unsupervised Feature Learning Framework for No-reference Image Quality Assessment. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1098–1105, 2012.