Displays 44 (2016) 21–26

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

Pairwise comparison and rank learning for image quality assessment

Long Xu [a,*], Jia Li [b], Weisi Lin [c], Yun Zhang [d], Yongbing Zhang [e], Yihua Yan [a]

[a] Key Laboratory of Solar Activity, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China
[b] State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[c] Nanyang Technological University, Singapore 639798, Singapore
[d] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[e] Graduate School at Shenzhen, Tsinghua University, Shenzhen 518005, China

ARTICLE INFO

Article history:
Received 19 January 2016
Accepted 20 June 2016
Available online 21 June 2016

Keywords:
Image quality assessment
Machine learning
Rank learning
Pairwise comparison

ABSTRACT

To know what kinds of image features are crucial for image quality assessment (IQA) and how these features affect the human visual system (HVS) is still largely beyond human knowledge. Hence, machine learning (ML) is employed to build IQA by simulating the HVS behavior in IQA processes. Support vector machine/regression (SVM/SVR) is a major member of ML. It has been successfully applied to IQA recently. As to image quality rating, the human's opinion about it is not always reliable. In fact, the subjects cannot precisely rate the small difference of image quality in subjective testing, resulting in unreliable Mean Opinion Scores (MOSs). However, they can easily identify the better/worse one from two given images, even their qualities do not differ much. In this sense, the human's opinion on pairwise comparison (PC) of image quality is more reliable than image quality rating. Thus, PC has been exploited in developing IQA metrics. In this paper, a rank learning optimization framework is firstly developed to model IQA. Particularly, the PCs of image quality instead of numerical ratings are incorporated into the optimization framework. Then, a novel no-reference (NR)-IQA is proposed to infer image quality in terms of image quality ranks. By importing rank learning theory and PC into IQA, a fundamental and meaningful departure from the existing framework of IQA could be expected. The experimental results confirm that the proposed Pairwise Rank Learning based Image Quality Metric (PRLIQM) can achieve comparable performance over the state-of-the-art NR-IQA approaches.

1. Introduction

There have been dramatically increased interest in image quality assessment (IQA) recently. The most accurate and reliable way of IQA is to ask the subjects who are shown a group of images for their opinions about the quality of these images. This way called subjective IQA, however, is highly time-consuming, human labor consuming, and impractical in real-time application. Thus, a plenty of objective IQA approaches have been developed during last decade. Based on the availability of references, IQA approaches can be classified into full-reference (FR), no-reference (NR) and reduced-reference (RR) approaches. In FR category, structural similarity (SSIM) [1] has been investigated extensively by the researchers due to its simple philosophy and mathematical form, as well as good performance.

Concerning real-world application, NR approaches are more general and applicable than FR approaches. We categorize the NR approaches of the literatures into three categories. The ***first*** one analyzes the behavior of specific distortion for IQA. In [2], Sheikh et al. employed wavelet statistical model to capture JPEG compression distortion. Liang et al. [3] combined the sharpness, blurring, and ringing measurements together to evaluate images distorted by JPEG 2000. In [4], Ferzli et al. introduced just noticeable blur into probability summation model to measure sharpness/blurriness. In [5], Brandao et al. exploited the DCT statistics of JPEG compression to establish a NR-IQA approach for assessing quality of images coded by JPEG. The ***second*** one uses quality aware clustering which arranges image patches of training set into several clusters according to certain local image features, such as histogram of oriented gradients (HoG), difference of Gaussian (DoG) and Gabor filter. Each cluster centroid is assigned quality by averaging the qualities of image patches in this cluster. By associating cluster centroid with its quality, codebook can be established. It performs like a dictionary. Each time, given a new image patch, we look up codebook to find the mostly matched codeword, and then retrieve

* Corresponding author.
E-mail addresses: lxu@nao.cas.cn (L. Xu), jiali@buaa.edu.cn (J. Li), WSLin@ntu.edu.sg (W. Lin), yun.zhang@siat.ac.cn (Y. Zhang), zhang.yongbing@sz.tsinghua.edu.cn (Y. Zhang), yyh@nao.cas.cn (Y. Yan).

http://dx.doi.org/10.1016/j.displa.2016.06.002
0141-9382/

its corresponding quality. In [6], visual codebook is formed on Gabor filter based local appearance descriptors. In [7], Wu et al. used FSIM [8] to compute quality of image patch instead of MOS to establish codebook. The ***third*** one utilizes machine learning (ML) tools, such as support vector machine/regression (SVM/SVR), Adaboost and Clustering [9–12], to map image features onto image quality ratings. In [13], Moorthy et al. employed SVM and SVR to learn a classifier and an ensemble of regressors for distortion classification and quality rating respectively. In [14], Tang et al. proposed an approach similar to [13] but with more elaborate features, including distortion texture statistics, blur/noise statistics and histogram of each subbands of image decomposition.

In [13,14], the distance between MOS and predicted image quality was optimized. Such an optimization objective cannot address IQA very well for the reasons: (1) *the numerical image quality, e.g., with rate of 1–5, is not exactly with a strong confidence for measuring real image quality. The small difference of image quality ratings may not truly reflect the real difference of image qualities*; (2) *to assess image quality, pairwise competition is more reliable/reasonable than numerical quality rating. The subjects are only requested to indicate the binary opinion (better or worse) to two compared images. This kind of comparison is less taxing and confusing than numerical rating system*; (3) *the diversity of image content and distortion types also make it difficult to rate image quality numerically under complex scenarios, but pairwise comparison (PC) is not that difficult*. To address these issues, PC of image quality has been introduced into IQA for assisting image quality rating. Since PC concerns $n \times (n-1)/2$ times of comparisons given n images, it is very labor consuming for acquiring MOSs in subjective experiment.

Two related works have been reported in [19,20]. In [19], image quality preference in pairs were exploited to lead to a rank learning optimization problem, and SVM with multiple kernel were adopted to solve this optimization problem. In [20], an approach was developed for ranking image enhanced algorithms, where image quality ranking rather than giving physical quantity of image quality was investigated. Both [19,20] were associated with a pairwise rank learning (PRL) [15,16] framework. Since PC of image quality only concerns binary option of image quality competition, PRL optimizations were realized by a binary classifier in both [19,20], and SVM/SVR was employed to do classification.

In this work, PC of image quality is formulated into a new PRL framework [21] which was originally used for saliency model. This framework forms PRL task as a general optimization problem instead of a binary classifier as mentioned above. In addition, it uses steepest descent method to solve this optimization problem, which would be faster than SVM, so it would be suitable for large-scale database. Moreover, we additionally take the quality difference intensity into consideration besides binary competition (better or worse) of image quality by introducing scaling factors which account for image quality difference of each pair of images.

The rest of this paper is organized as follows. Section 2 describes the proposed PRLIQM in detail. Section 3 presents the experimental results. And, the final section concludes this paper.

2. Proposed pairwise rank learning based image quality metric

Recently, there is a new trend to establish NR-IQA models by using ML [6,13,14,22–25]. Inspired by the development of rank learning in information retrieval (IR) [15–18], we make a fundamental departure from the family of existing ML based approaches. And, a new PRL framework is proposed with two distinct characteristics from previous ones: (1) *it is established on a rank learning framework*; (2) *only logical comparison instead of numerical rating of image quality is concerned.* The proposed PRL only requires the variable of MOS to be ordinal, while the conventional ML based approaches need an assumption of interval variable for MOS since the numerical computing and statistics are used there (please refer to [26] for the definitions of "ordinal" and "interval"). Therefore, it is more applicable in real-world applications.

Regarding rank learning, the deduced computer model targets at ranking objects instead of assigning a physical quantity of image quality (like PSNR) to each object. Usually, in IR, it ranks the retrieved items by their relevance with the query. To our concerned IQA, we measure image qualities by their ranks instead of physical quantities. Thus, the computer model derived from rank learning rank images firstly. Then, a relation between MOSs and ranks can be established by using polynomial curve fitting. In addition, the pairwise approach as stated in [26] is employed to establish optimization objective function, where the binary comparisons of MOSs are to be ground-truth for training computer model, and the risk function [26] is based on indicator (0–1) loss function which has the binary outputs of 0 and 1, representing inconsistence and consistence between predicted rank of image quality and ground-truth respectively. The related issues of rank learning based optimization are to be detailed in this section.

2.1. Training data for rank learning

We carry out our work on image quality rating databases, such as LIVE image database [27], with numerical ratings of image qualities, i.e., MOSs given by subjects. For conventional ML based training task, we assume the feature vectors $\{x_i\}(i = 1, 2, \ldots, n)$, and labels $\{y_l\}(l = 1, 2, \ldots, k)$ given by MOS. Generally, feature vector concerns high level information of a visual scene, which is extracted from image by using some local/global image descriptors, such as difference of Gaussian (DoG), Gabor filter, wavelet coefficients, Fourier coefficients and newly developed deep learning techniques [28].

To establish the pairwise rank learning task on image quality rating system for IQA, the ground-truth is given by comparing images in pair with respect to their MOSs. Given MOSs $\{y_i\},\ i = 1 \ldots n$, a binary label $\{+1, -1\}$ is assigned to $y_i \geqslant y_j$ and $y_i < y_j$ respectively.

2.2. Pairwise rank learning model

SVM is a supervised learning tool of ML category. The objective of SVM is a little sophisticated relative to general optimization objective of least square error. It optimizes the maximum margin between two classes of samples. There are some variants of SVM, such as L1-SVM, L2-SVM and least squares (LS) SVM. We explore the intrinsic principle of ML for IQA, by optimizing the numerical distance between predicted image quality ($\varphi_\omega(x_i)$) and MOS (y_i) as

$$\begin{aligned} \omega^* &= \arg\min_{\omega} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \hat{\xi}_i), \\ s.t.\quad & y_i - \varphi_\omega(x_i) \leqslant \varepsilon + \xi_i, \forall i \\ & \varphi_\omega(x_i) - y_i \leqslant \varepsilon + \hat{\xi}_i, \forall i \\ & \xi_i \geqslant 0, \hat{\xi}_i \geqslant 0 \end{aligned} \tag{1}$$

where φ_ω is a model parameter learned by resolving (1), and used to compute image quality for unknown input image; x_i represents image features of the i-th image, y_i is the label of x_i given by MOS, and $\|\cdot\|_p$ represents p-norm operation. The linear form of $\varphi_\omega:\ \varphi_\omega = \omega^T x$, is widely used in the literature. For fitting MOS curves more generally, nonlinear functions are employed, which explore the nonlinear relationship between image features and MOS. By using kernel functions, nonlinear problems can be converted into linear problems. Observing the optimization objective of (1), the p-norm is optimized, while a new optimization objective

is based on binary comparison of image quality is established in this work as

$$\min_{\omega}\left\{\sum_{i\neq j}[y_j-y_i]_+[\varphi(x_i)\geqslant\varphi(x_j)]_I\right\}, \tag{2}$$

where $[x]_I=1$ if the logic decision x holds; otherwise $[x]_I=0$; $[x]_+=\max\{0,x\}$. Eq. (2) is constructed on the ranks of image qualities instead of the numerical values. From (2), if two images ranks wrong, i.e., their ranks violate the ground-truth given by MOSs, the cost of (2) would increase. In addition, (2) is cost-sensitive since the operation $[x]_+$ contributes a weight to the summation of (2). An image with large quality difference from others would contribute more to (2), while the images with similar image qualities tend to have low weights to the cost $L(\omega)$. This phenomena just coincides with our statement about drawbacks of the optimization on image quality rating. A concept map is drawn in Fig. 1 for explaining this phenomena. From Fig. 1, if an image is far from others regarding image quality, it would have a higher weight to the optimization objective. For example, the samples labeled by red circle should be crucial to train a computer model from (2), however the p-norm optimization objective of (1) ignores these samples since the overwhelming majority of the samples (95%) are in a straight line. In rank based regression, each sample in the subset in red circle would compare with all samples in other two subsets. Therefore, these 5% samples are crucial in optimization although they only account for a small percent of the training set.

For simplicity, the linear function $\varphi(x)=\varphi^Tx$ is assumed in (2). Thus, the optimization objective is to seek a vector ω which results in the minimum of (2) on the training set. With a linear function $\varphi(x)$, (2) is rewritten as

$$\min_{\omega}\left\{\sum_{i\neq j}[y_j-y_i]_+[\omega^Tx_i\geqslant\omega^Tx_j]_I\right\}. \tag{3}$$

Let $L(\omega)=\sum_{i\neq j}[y_j-y_i]_+[\omega^Tx_i\geqslant\omega^Tx_j]_I$, we call $L(\omega)$ the empirical loss. Since $[x]_I$ is non-convex, we encounter a non-convex optimization problem. As in [21,29], the Boolean terms related to ω in (3) is replaced by their upper bounds to facilitate the optimization as

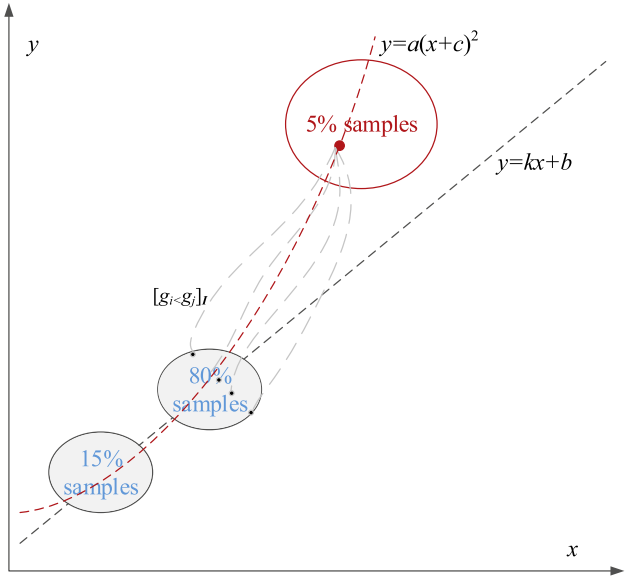

Fig. 1. Illustrating the significance of the 5% percent of samples in red circle to optimization of (2) (however, this part of samples play insignificantly in L_2 norm optimization of (1)).

$$[\omega^Tx_i\geqslant\omega^Tx_j]_I\leqslant e^{(\omega^Tx_i-\omega^Tx_j)}, \tag{4}$$

where the exponential upper bound is used since it is convex and can facilitate the optimization. After the only one term containing the variable ω in (3) is replaced, the empirical loss function would turn out to be convex. Then, the gradient decedent method can be employed to solve (3). Note that we have

$$\frac{\partial}{\partial\omega}e^{(\omega^Tx_i-\omega^Tx_j)}=(x_i-x_j)e^{(\omega^Tx_i-\omega^Tx_j)}, \tag{5}$$

so the gradient decent direction can be written as

$$\Delta\omega=\lambda\times\sum_{i\neq j}[y_j-y_i]_+(x_i-x_j)e^{(\omega^Tx_i-\omega^Tx_j)}, \tag{6}$$

where λ acts as an iteration step controlling the convergence speed.

From (3), given $\{y_i\}$, $\{x_i\}$ and an initial ω, the empirical loss $L(\omega)$ can be initialized. Replacing ω by $\omega+\Delta\omega$, $L(\omega)$ can be updated. By iteratively updating ω and $L(\omega)$, the global minimum objective can be reached.

2.3. Mapping image quality rank to image quality score

In PRL framework, the optimization objective function (3) is established on PC of image qualities instead of numerical image quality ratings. Thus, only image quality ranks can be provided for testing. To give image quality to a test image, it is needed to convert image quality ranks to image quality ratings/scores. In [19], the number of times of that an image is preferable against the others was defined as "gain" which is proportional to the perceived quality of that image. In addition, a linear mapping between the gain and the quality score was assumed and fitted by training data. Thus, after computing image quality ranks, their quality scores can be deduced by the two steps mentioned above. However, [19] needs the difference between the test image and each training image which is unavailable in real application. In [20], only competition of image enhancement algorithms was concerned without the need of conversion from image quality rank to physical quantity of image quality. The proposed PRLIQM can output a rank list of all images instead of only binary preference of each two images as in [19,20]. Therefore, a mapping function between image quality ranks and scores can be deduced from PRLIQM directly as follows. In training stage of PRLIQM, a nonlinear fitting function is deduced from mapping the predicted rank list to MOSs rank list. In test stage, the nonlinear fitting function can tell image quality score of each test image without the need of any information from training set.

The i-th image is compared with other images in training set. We count the number of wrong ranks, i.e., the predicted ranks is inverse of ground-truth, and compute the accumulative image quality difference (AIQD) as

$$g(i)=\frac{\sum_{i\neq j}^{n}[\omega^Tx_i-\omega^Tx_j]_+}{\sum_{i\neq j}^{n}[\omega^Tx_i>\omega^Tx_j]_I}. \tag{7}$$

Since $[\omega^Tx_i-\omega^Tx_j]_+$ indicates the quality of the i-th image relative to the j-th image, Eq. (7) can represent the relative quality of the i-th image against others. We draw the relationship between $g(i)$ and MOSs in Fig. 2. A nice fitting curve can be observed from Fig. 2. In addition, the shape of this mapping function is well fit by an exponential function as

$$q(i)=\beta_1+\beta_2\times\exp\{\beta_3\times g(i)\} \tag{8}$$

The parameters can be easily deduced from nonlinear least squares regression (implemented by "*nlinfit*" function in Matlab). One can derive image quality score given image quality rank from (8). The advantage of (8) lies in that accumulative summation reduce the

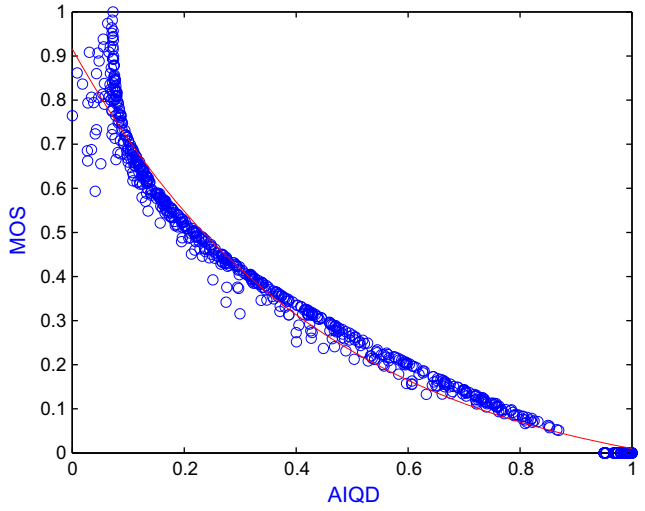

Fig. 2. Mapping between relative quality AIQD ($g(i)$) and MOSs: AIQD is computed from (7) represents relative quality; the vertical axis represents image quality given by MOSs; the curve in red color is the fitting function (in this case, $\beta_1 = -0.0949$, $\beta_2 = 1.0113$ and $\beta_3 = -2.2605$) by using nonlinear least squares regression. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

interference of noise, so a good fitting curve is deduced as shown in Fig. 2. Observing (3), PRLIQM is based on accumulative summation, so AIQD is reasonable and fit for representing relative image quality. The experiments performed in Section 3 proves its good performance.

3. Experimental results and discussions

3.1. Databases and evaluation protocols

The performance of an IQA metric can be evaluated by depicting the correlation between predicted image qualities and subjective image qualities. The MOSs are the ground-truth for evaluation of IQA. We perform the experiments on LIVE image database [27], which consists of 29 reference images, each image has 5 distortion types (JPEG, JP2K, white noise (WN), Gaussian blur (GB) and fast fading (FF) channel distortions) and 5/6 distortion levels per type. The images in database are divided into training set and testing set. A training set consists of 80% of the images in the database, and a testing set consists of the remaining 20% of the images. In order to ensure that PRLIQM is robust across content and is not biased by the specific train-test split, random 80% train-20% test split is repeated 1000 times on database. Each time uses the different training-testing split. This configuration of train-test split is the same as the configurations of [6,13,14,22–25].

To compete with the state-of-the-art metrics, the conventional correlation measurements are compared in this subsection. Three parameters: Pearson's linear correlation coefficient (PLCC), root mean square error (RMSE), and Spearman's rank order correlation coefficient (SROCC), are used to measure correlation. The PLCC between two data sets, A and B, is defined as

$$PLCC(A,B) = \frac{\sum_{i=1}^{n}(A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^{n}(A_i - \bar{A})^2}\sqrt{\sum_{i=1}^{n}(B_i - \bar{B})^2}}, \quad (9)$$

which measures the linear correlation coefficient between A and B, so it can represent the prediction accuracy of B to A. SROCC is computed on the ranked A and B with respect to their magnitudes, so it can evaluate the prediction monotonicity, i.e., the degree of the predictions of a metric agree with the relative magnitudes of MOSs. RMSE measures the error during fitting process. Larger PLCC and SROCC values indicate better correlation between objective image quality scores and MOSs, while smaller RMSE values mean smaller error of predictions, therefore a better performance.

It is a crucial step to extract image features (feature vector) being the representation of an image in ML methods. In this work, the feature vector is given by referring to natural statistic model (NSS) model [30] which has been widely used in the state-of-the-art ML based methods [13,22–25]. NSS model acquires low-level statistical properties of images to measure the destruction of naturalness caused by distortions.

3.2. Performance on individual database

As mentioned in Section 2, PRLIQM outputs image quality ranks. We need to convert these ranks into image quality scores as explained in Section 2.3. Following (8), the relationship between AIQD and MOSs is drawn for test images in Fig. 3. It can be observed that the curve shape on test set is highly similar to Fig. 2, which implies that such a mapping is consistent between training and testing. After converting image quality ranks into image quality numbers, the media PLCC, SROCC and RMSE values across 1000 times of training-testing split are tabulated in Tables 1 and 2, for each distortion category, as well as across distortion categories.

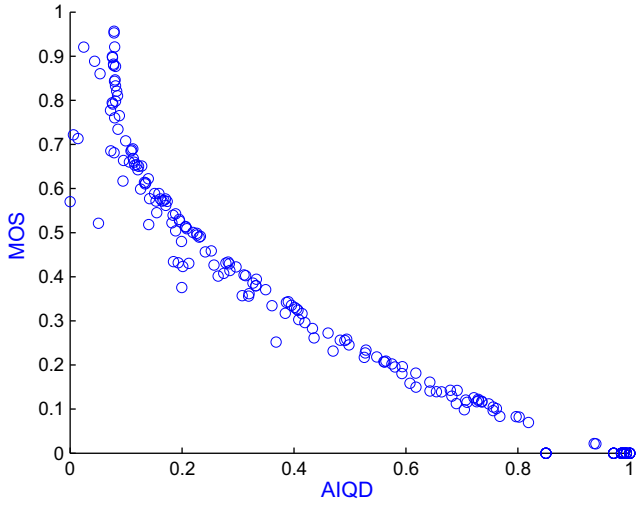

Fig. 3. Mapping between relative quality AIQD and MOSs on test set: AIQD is computed from (7) represents relative quality; the vertical axis represents image quality given by MOSs.

Table 1
Median PLCC across 1000 train-test combinations on **live image** database.

	JP2K	JPEG	WN	Blur	FF	All
PSNR	0.8762	0.9029	0.9173	0.7801	0.8795	0.8592
SSIM	0.9405	0.9462	0.9824	0.9004	0.9514	0.9066
MS-SSIM	0.9746	0.9793	0.9883	0.9645	0.9488	0.9511
VIF	0.9790	0.9880	0.9920	0.9760	0.9720	0.9610
CBIQ	0.8898	0.9454	0.9533	0.9338	0.8951	0.8955
LBIQ	0.9103	0.9345	0.9761	0.9104	0.8382	0.9087
BLIINDS-II	0.9386	0.9426	0.9635	0.8994	0.8790	0.9164
DIIVINE	0.9233	0.9347	**0.9867**	0.9370	0.8916	0.9270
BRISQUE	0.9229	**0.9734**	**0.9851**	0.9506	0.9030	**0.9424**
TMIQ	0.8730	0.8941	0.8816	0.8530	0.8234	0.7856
NIQE	0.9370	0.9564	0.9773	**0.9525**	0.9128	0.9147
BIQA	–	–	–	–	–	–
PRLIQM-I	**0.9406**	0.9416	0.9494	0.9446	**0.9559**	0.9403
PRLIQM-II	**0.9443**	**0.9575**	0.9553	**0.9641**	**0.9533**	**0.9606**

Table 2
Median SROCC across 1000 train-test combinations on **live image** database.

	JP2K	JPEG	WN	Blur	FF	All
PSNR	0.8646	0.8831	0.9410	0.7515	0.8736	0.8636
SSIM	0.9389	0.9466	0.9635	0.9046	0.9393	0.9129
MS-SSIM	0.9627	0.9785	0.9773	0.9542	0.9386	0.9535
VIF	0.9670	0.9820	0.9840	0.9730	0.9630	0.9640
CBIQ	0.8935	0.9418	0.9582	0.9324	0.8727	0.8954
LBIQ	0.9040	0.9291	0.9702	0.8983	0.8222	0.9063
BLIINDS-II	0.9323	0.9331	0.9463	0.8912	0.8519	0.9124
DIIVINE	0.9123	0.9208	**0.9818**	0.9373	0.8694	0.9250
BRISQUE	0.9139	**0.9647**	**0.9786**	**0.9511**	0.8768	**0.9395**
TMIQ	0.8412	0.8734	0.8445	0.8712	0.7656	0.8010
NIQE	0.9172	0.9382	0.9662	0.9341	0.8594	0.9135
BIQA	**0.9440**	0.9450	0.9730	0.9530	0.9080	0.9380
PRLIQM-I	0.9323	0.9336	0.9351	0.9408	**0.9432**	0.9326
PRLIQM-II	**0.9493**	**0.9546**	0.9528	**0.9651**	**0.9417**	**0.9488**

In Tables 1 and 2, the top two metrics are highlighted in bold font. As can be seen from Tables 1 and 2, PRLIQM-II is with the best or second-best performance among all compared NR-IQA metrics with regard to PLCC and SROCC which indicate it can predict image quality objectively with the considerable good performance. Remarkably, PRLIQM-I is much better than DIIVINE. Since they use the same feature, the superiority of PRLIQM-I only comes from the pairwise rank learning. In addition, PRLIQM-II is better than PRLIQM-I. It uses a new feature consists of NSS and additional 96 entries representing histogram of subbands of PD.

3.3. Performance across databases

To verify the proposed method is independent of database, we train the model on the entire LIVE database (all of the five distortion types), and test it on the CSIQ database. We report the PLCC and SROCC statistics on each distortion type and across all distortion types in Tables 3 and 4. Here, we only provide the comparison among the NR-IQA metrics since database dependent does not exist in FR-IQA. The bold fonts highlight the top two metrics. It can be observed that PRLIQM-II always ranks in the top two metrics. Since the additive Gaussian pink noise ("PN") and global contrast decrements ("GCD") distortion types are not included in LIVE database for training, so they are excluded from testing, and only "JPEG", "JPEG2000", White noise ("WN") and Gaussian blur ("Gblur") are tested and listed in Tables 3 and 4. It can be observed that PRLIQM can predict image quality well on individual

Table 3
PLCC comparison on **CSIQ** database.

	WN	JPEG	JP2k	Gblur	AllSub
DIVINE	**0.8825**	0.8834	0.8836	0.8823	0.6573
BLIINDS-II	0.7984	0.8311	0.8146	0.8100	0.6727
BRISQUE	0.7148	0.8118	0.8389	**0.9183**	0.8005
NIQE	0.8115	**0.9344**	**0.9260**	**0.9263**	**0.8871**
PRLIQM-I	0.8798	0.9014	0.9161	0.9021	0.8732
PRLIQM-II	**0.8836**	**0.9115**	**0.9251**	0.9056	**0.9073**

Table 4
SROCC on **CSIQ** database.

	WN	JPEG	JP2k	Gblur	AllSub
DIVINE	**0.8662**	0.8689	0.8692	0.8667	0.5621
BLIINDS-II	0.8009	0.8413	0.8153	0.7914	0.5999
BRISQUE	0.6678	0.7838	0.8047	0.8785	0.7467
NIQE	0.8097	0.8821	**0.9063**	**0.8943**	**0.8705**
BIQA	0.8240	0.8570	**0.8850**	0.8450	0.8480
PRLIQM-I	0.8565	**0.8874**	0.8721	0.8862	0.8636
PRLIQM-II	**0.8689**	**0.9019**	0.8842	**0.9025**	**0.8721**

Table 5
Feature complexity comparisons (average processing time (in second) per image).

	PSNR	SSIM	MS-SSIM	EPSNR	ESSIM	FSIM	VIF	TMIQ	CBIQ	LBIQ	DIIVINE	BLIINDS-II	NIQE	BRISQUE	PRLIQM-I	PRLIQM-II
Processing time	0.0258	0.05	0.0743	0.0454	0.0880	0.2735	0.99	36.8599	1.2724	–	17.8998	59.8810	0.2729	0.2020	18.0213	18.2422
Feature length	N/A	N/A	N/A	N/A	N/A	N/A	N/A	36/block	40/block	16689/image	88/image	83/image	182/image	182/image	88/image	184/image

distortion type and across distortion types. PRLIQM-II is the best among all benchmarks on almost all individual distortion types; and it has the highest PLCC and SORCC correlations across distortion types of "JPEG", "JPEG2000", "WN" and "Gblur". In Tables 3 and 4, "AllSub" represents that the correlations are computed on these four distortion types excluding "PN" and "GCD". This experiment proves that the proposed PRLIQM is independent of database, so it can be employed in many actual application.

3.4. Feature complexity

PD was performed for computing NSS in [13]. The same configuration as [13] is employed to compute NSS in this work. There are 6 scales and 2 directions. For PRLIQM-II, the normalized histogram of each scale is additionally computed to contaminate with the original NSS feature to output a new one. In addition, the steepest gradient descent (SGD) method is employed to resolve the proposed optimization function, which has less computational complexity than ML-based methods.

The computational complexity is compared among PRLIQM and benchmarks with respect to computing time. The statistics of computing time are listed in Table 5. The computer configuration: *Intel (R) Xeon (R) CPU 3.1 GHz (2 CPUs, 4 threads); 16G 1600 MHz RAM; Windows 7 professional OS and Matlab2013b*. From Table 5, PRLIQM is comparable to DIIVINE, and inferior to LBIQ, CBIQ, NIQE and TMIQ regarding computational complexity. TMIQ is the most complex among all methods. Here, the computing time only concerns the implemented time of IQA process, excluding the training process which is offline regarding IQA process. In training process of PRLIQM, the number of interactions is about 25 at average. Since SGD is used to solve optimization, the implementation is very fast.

We regard feature complexity from both computational complexity and dimension/length of features. Since the small number of samples in subjective quality database, too large dimension of features tends to arouse overfitting. We compare the dimension of features among the proposed and benchmarks in Table 5. PRLIQM has 88 and 184 length of feature vector, and the former is same with DIIVINE [13]. LBIQ has the largest dimension of features, so principle component analysis (PCA) is required to compress the dimension of features before training process. Thus, PRLIQM is competitive among all testing algorithms with respect to feature complexity. Remarkably, the features of PRLIQM-I and DIIVINE are the same in Table 5, so the achievement of PRLIQM over DIIVINE shown in Tables 1 and 2 is only obtained from pairwise rank learning.

4. Conclusions

In this paper, we have investigated PRL for IQA, in which the image quality ranking instead of the numerical image quality rating is exploited to build IQA models. Since PC is less taxing and confusing than image quality rating, especially for images with small quality difference, it could be complementary to the current image quality rating system. Therefore, the proposed PRL would be promising for developing IQA metrics. In addition, PRL is preferable in the situations that the measures themselves are of less interest than their ranks, so it could be beneficial to other optimizations concerning PC potentially. Last but not least, the efforts (including ours) on PRL make a fundamental departure from existing studies on IQA, which may raise a new perspective of IQA in the near future.

References

[1] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (2004) 600–612.
[2] H.R. Sheikh, A.C. Bovik, L. Cormack, No-reference quality assessment using nature scene statistics: JPEG 2000, IEEE Trans. Image Process. 14 (2005) 1918–1927.
[3] L. Liang, S. Wang, J. Chen, S. Ma, D. Zhao, W. Gao, No-reference perceptual image quality metric using gradient profiles for JPEG 2000, Signal Process.: Image Commun. 25 (2010) 502–516.
[4] R. Ferzli, L.J. Karam, A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB), IEEE Trans. Image Process. 18 (2009).
[5] T. Brandao, M.P. Queluz, No-reference image quality assessment based on DCT domain statistics, Signal Process. 88 (2008) 822–833.
[6] P. Ye, D. Doermann, No-reference image quality assessment using visual codebooks, IEEE Trans. Image Process. 21 (2012) 3129–3138.
[7] W. Xue, L. Zhang, X. Mou, Without human scores for blind image quality assessment, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 995–1002.
[8] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: a feature similarity index for image quality assessment, IEEE Trans. Image Process. 20 (2011) 2378–2386.
[9] B. Gu, V.S. Sheng, K.Y. Tay, W. Romano, S. Li, Incremental support vector learning for ordinal regression, IEEE Trans. Neural Netw. Learn. Syst. 26 (2015) 1403–1416.
[10] B. Gu, V.S. Sheng, Z. Wang, D. Ho, S. Osman, S. Li, Incremental learning for γ-Support vector regression, Neural Netw. 67 (2015) 140–150.
[11] X. Wen, L. Shao, Y. Xue, W. Fang, A rapid learning algorithm for vehicle classification, Inf. Sci. 295 (2015) 395–406.
[12] Y. Zheng, B. Jeon, D. Xu, et al., Image segmentation by generalized hierarchical fuzzy C-means algorithm, J. Intell. Fuzzy Syst. 28 (2015) 961–973.
[13] A.K. Moorthy, A.C. Bovik, Blind image quality assessment: from natural scene statistics to perceptual quality, IEEE Trans. Image Process. 20 (2011) 3350–3364.
[14] H. Tang, N. Joshi, A. Kapoor, Learning a blind measure of perceptual image quality, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 305–312.
[15] H. Li, Learning to rank for information retrieval and natural language processing, Synth. Lect. Human Lang. Technol. 4 (2011) 1–113.
[16] T. Liu, Learning to rank for information retrieval, Found. Trends Inform. Ret. 3 (2009) 225–331.
[17] Z. Xia, X. Wang, X. Sun, Q. Wang, A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data, IEEE Trans. Parallel Distrib. Syst. 27 (2016) 340–352.
[18] Z. Fu, X. Sun, Q. Liu, L. Zhou, J. Shu, Achieving efficient cloud search services: multi-keyword ranked search over encrypted cloud data supporting parallel computing, IEICE Trans. Commun. E98-B (2015) 190–200.
[19] F. Gao, D. Tao, X. Gao, X. Li, Learning to rank for blind image quality assessment, IEEE Trans. Neural Netw. Learn. Syst. 26 (2015) 2275–2290.
[20] Z.Y. Chen, T.T. Jiang, Y.H. Tian, Quality assessment for comparing image enhancement algorithms, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3003–3010.
[21] J. Li, Y.H. Tian, T.J. Huang, W. Gao, Cost-sensitive rank learning from positive and unlabeled data for visual saliency estimation, IEEE Signal Process. Lett. 17 (2010) 591–594.
[22] M. Saad, A.C. Bovik, C. Charrier, Blind image quality assessment: a natural scene statistics approach in the DCT domain, IEEE Trans. Image Process. 21 (2012) 3339–3352.
[23] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (2012) 4695–4708.
[24] A. Mittal, G.S. Muralidhar, J. Ghosh, A.C. Bovik, Blind image quality assessment without human training using latent quality factors, IEEE Signal Process. Lett. 19 (2011) 75–78.
[25] A. Mittal, R. Soundararajan, A.C. Bovik, Making a "completely blind image quality analyzer, IEEE Signal Process. Lett. 20 (2013) 209–212.
[26] L. Xu, J. Li, W. Lin, Y. Zhang, L. Ma, Y. Fang, Y. Yan, Multi-task rank learning for image quality assessment, IEEE Trans. Circ. Syst. Video Technol. 2016 (in press), http://dx.doi.org/10.1109/TCSVT.2016.2543099.
[27] LIVE Image Quality Database. Available [Online]: <http://live.ece.utexas.edu/research/quality/subjective.htm>.
[28] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1798–1828.
[29] L. Xu, W.S. Lin, J. Li, Y.M. Fang, Y.H. Yan, Rank learning on training set selection and quality assessment, in: Proc. IEEE International Conference on Multimedia and Expo (ICME), 2014, pp. 1–6.
[30] Z. Wang, A.C. Bovik, Reduced- and no-reference image quality assessment, IEEE Signal Process. Mag. 28 (2011) 29–40.