Optimizing Multi-Stage Discriminative Dictionaries for Blind Image Quality Assessment

Qiuping Jiang, *Student Member, IEEE,* Feng Shao, *Member, IEEE,* Weisi Lin, *Fellow, IEEE,* Ke Gu, Gangyi Jiang, *Member, IEEE,* and Huifang Sun, *Fellow, IEEE*

***Abstract*—State-of-the-art algorithms for blind image quality assessment (BIQA) typically have two categories. The first category utilizes handcrafted natural scene statistics (NSS) derived from the statistical regularity of natural images. The second category utilizes codebook-based features which are obtained by feature encoding over a learned codebook. However, several problems need to be addressed in existing codebook-based BIQA methods. First, the high-dimensional codebook-based features are memory-consuming and have the risk of over-fitting. Second, there is a semantic gap between the constructed codebook by unsupervised learning and image quality. To address these problems, we propose a novel codebook-based BIQA method by optimizing multi-stage discriminative dictionaries (MSDDs). To be specific, MSDDs are learned by performing the label consistent K-SVD (LC-KSVD) algorithm in a stage-by-stage manner. For each stage, a new quality consistency constraint called "quality-discriminative regularization" term is introduced and incorporated into the reconstruction error term to form a unified objective function which can be effectively solved by LC-KSVD for discriminative dictionary learning. Then, the latter stage takes the reconstruction residual data in the former stage as input based on which LC-KSVD is repeatedly performed until the final stage is reached. Once the MSDDs are learned, multi-stage feature encoding (MSFE) is performed to extract feature codes. Finally, the feature codes are concatenated across all stages and aggregated over the entire image for quality prediction via regression. The proposed method has been evaluated on five databases and experimental results well confirm its superiority over existing relevant BIQA methods.**

***Index Terms*—Blind image quality assessment, multi-stage discriminative dictionaries, multi-stage feature encoding, label consistent K-SVD, reconstruction residual.**

I. Introduction

Manuscript received March 7, 2016; revised June 26, 2017 and September 2, 2017; accepted October 1, 2017. This work was supported in part by the Natural Science Foundation of China (61622109), in part by the Zhejiang Natural Science Foundation (R18F010008), and the China Scholarship Council (201708330302). It is also sponsored by the K.C. Wong Magna Fund in Ningbo University. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Judith Redi. *(Corresponding author: Feng Shao.)*

Q. Jiang is with the Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China, and also with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: jqp910707@126.com).

F. Shao and G. Jiang are with the Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China (e-mail: shaofeng@nbu.edu.cn; jianggangyi@nbu.edu.cn).

W. Lin is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wslin@ntu.edu.sg).

K. Gu is with the Beijing Key Laboratory of Computational Intelligence and Intelligent System, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China (e-mail: guke@bjut.edu.cn).

H. Sun is with the Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139 USA (e-mail: hsun@merl.com).

VISUAL data, especially in the form of digital images, has become omnipresent in this digital media era. There is a saying that one high-quality image is worth thousands of words. While it is true, it needs to be aware that digital images are likely to undergo a variety of distortions during the process chain from capture to display [1], [2]. Therefore, quantifying the impacts of distortions on image quality in a perceptually consistent way is important. The development in this direction will advance a wide range of applications, such as video coding [3], [4], [5], image restoration [6], image compression [7], etc.

Image quality assessment (IQA) metrics can be classified into full-reference (FR), reduced-reference (RR), and no-reference (NR) categories [8]. For FR-IQA metrics, the original images are fully available. Many effective FR-IQA metrics, including SSIM [9], MAD [10], ADM [11], GSM [12], FSIM [13], SFF [14], VSI [15], and GMSD [16], have been proposed. Owing to the full participation of original images, FR-IQA metrics have achieved high consistency with human subjective judgment results. Different from FR metrics, RR-IQA metrics [17], [18], [19], [20] only utilize partial original features for quality evaluation. As much less original information is involved, RR-IQA metrics are more efficient and practically applicable. However, in the case where the original image is unavailable, NR-IQA/blind IQA (BIQA) metrics would be the only possible solution. Note that, besides the conventional camera captured images, IQA algorithms for other images, such as stereoscopic image [21], [22], retargeted image [23], screen content image [24], and tone mapped image [25], [26], have also been investigated.

The initial efforts on BIQA mainly focus on specific distortion types, including blurriness [27], blockiness [28], and ringing artifacts [29]. The design of such distortion-specific BIQA models largely depends on domain knowledge of each specific distortion characteristics. Although satisfactory performance has been achieved, the universality of these methods is limited. To tackle this problem, BIQA is further developed to handle diverse distortions, known as the general-purpose BIQA. Much progress has been made in the areas related to general-purpose BIQA in recent years. According to the dependency of human opinion scores for the design of quality prediction models, state-of-the-art BIQA methods can be roughly classified as distance-based and learning-based.

The distance-based BIQA methods do not require human opinion scores to calibrate a quality model. They measure the image quality by the distance between the statistical models built upon pristine image sets and a testing distorted image.

Generally, these methods like [30], [31], [32] share a similar architecture. First, perceptually relevant features are extracted from pristine images and used to build a pristine statistical model. Then, the corresponding features are also extracted from a testing distorted image to build a distorted statistical model. Finally, the quality of this testing distorted image is estimated by the distance between the built two statistical models. In [30], the Natural Image Quality Evaluator (NIQE) extracts a set of local features and fits the feature vectors to a global multivariate Gaussian (MVG) model. Recently, the NIQE model is further extended to Integrated Local NIQE (IL-NIQE) [31] by enriching the quality-aware features and integrating the local and global MVG models together.

The learning-based BIQA methods require a set of distorted images with corresponding subjective scores to learn a quality model. The determinant factor leading to the success of learning-based BIQA methods is to extract highly versatile quality-aware features that are sensitive to a broad range of image distortion types while robust to image content variations. Broadly, there are two types of quality-aware features: natural scene statistic (NSS)-based and codebook-based. The NSS-based features are derived based on the assumption that natural images potentially possess certain regular statistical properties while the presence of distortions will inevitably change them. Many NSS-based methods in wavelet domain [33], DCT domain [34], spatial domain [35], and hybrid domain [36] have been proposed. Other works are also included in [37], [38], [39], [40].

In sharp contrast to the handcrafted NSS-based features, codebook-based features are obtained by feature encoding with respect to a codebook learned from raw patches. Generally, codebook-based features require less domain knowledge and are potentially generalizable across different image types [43]. The Bag-of-Words (BoW) model [45], which has been widely used in image classification [46], has also been adapted for BIQA [41], [42], [43], [44]. Given that BIQA is essentially a typical regression problem, the application of BoW to BIQA is intuitive. Typical codebook-based BIQA methods, including CBIQ [41], CORNIA [42], HOSA [43], and QAF [44], shares a similar architecture, i.e., local feature extraction, codebook construction, feature encoding, spatial pooling, and quality regression. With codebook-based feature representation, these methods have shown advantages in assessing both natural scene and screen content images [43].

It is worth noting that previous codebook-based methods [41], [42], [43], [44] usually require a large-size codebook to extract high-dimensional features, which are memory-consuming and have the risk of over-fitting. Additionally, codebooks used in previous methods are constructed by unsupervised learning, i.e., the quality information of training samples is not utilized during codebook optimization. Thus, the constructed codebooks may be suboptimal for BIQA. We interpret these two aspects of problems are related to some extent. On the one hand, due to the lack consideration of quality information of training samples during codebook optimization, it is difficult to produce highly discriminative feature codes with respect to a single small-size codebook because the codebook is not exclusively optimized to be quality-aware. Moreover, feature encoding over a single small-size codebook would inevitably cause large information loss and reconstruction error. On the other hand, by including sufficient (probably redundant) codewords in the codebook, a competitive performance still can be achieved by machine learning-based regression. From the above analyses, it is intuitive that optimizing discriminative (quality-aware) codebook for feature encoding and making use of the reconstruction residual data will facilitate the construction of codebook-based BIQA methods with a much smaller-size codebook.

This paper proposes a novel codebook-based BIQA method from the perspective of optimizing multi-stage discriminative dictionaries (MSDDs) for feature encoding to extract more compact and discriminative quality-aware features. To be specific, MSDDs are learned from a set of contrast normalized patches by performing label consistent K-SVD (LC-KSVD) [47] in a stage-by-stage manner. During each stage, a new quality consistency constraint called "quality-discriminative regularization" term is introduced and incorporated into the reconstruction error term to form a unified objective function which can be effectively solved by LC-KSVD. Then, the latter stage takes the reconstruction residual data in the former stage as input based on which LC-KSVD is repeatedly performed until the final stage is reached. Once the MSDDs are learned, multi-stage feature encoding (MSFE) is performed to extract feature codes. Finally, the feature codes are concatenated across all stages and aggregated over the entire image for quality prediction via support vector regression (SVR) [48]. Compared with the existing codebook-based BIQA methods, the proposed method extracts much lower dimensional features for quality evaluation while achieving comparable or even better performance.

The main advantages of our proposed method are two-fold. The first one is the high efficacy. Instead of performing feature encoding over a single large-size reconstructive codebook, MSDDs are learned by applying LC-KSVD in a stage-by-stage manner for MSFE to extract more compact and discriminative quality-aware features, thus the prediction accuracy is improved. The extracted features are shown to be applicable to natural images (degraded with single distortion type and multiple distortion types) and screen content images while the leading NSS-based model (e.g., GM-LOG [37]) can only effectively evaluate the natural images degraded with single distortion type as demonstrated by the experiments in Section IV. It should be emphasized that the proposed method is most suitable to evaluate the images degraded with some commonly encountered distortions such as JPEG2000 Compression (JP2K), JPEG Compression (JPEG), Gaussian White Noise (WN), Gaussian Blur (GB) and Fast Fading (FF). While for luminance change and color/contrast distortions, our method is still unable to deliver satisfactory results. However, given that these distortions are the primary challenge of state-of-the-art universal blind quality metrics, a slight performance boost on these distortions (as demonstrated by the experiments on the entire TID2013 database [49], see Section IV-E) achieved by our method is still acceptable. All these failed cases could lead us to develop more robust and universal BIQA models in the future by learning more comprehensive quality-aware

features. The second one is the high efficiency. The efficiency of an algorithm typically involves two factors: computational complexity and occupied memory. Compared to the existing codebook-based methods, much lower dimensional features are extracted in our method. Thus, our proposed method is more memory-saving, leading to high efficiency. The reduction in feature dimension is meaningful for the systems where the memory sources are limited. Although the computational complexity of our method is slightly higher than HOSA [43] which actually represents the state-of-the-art codebook-based BIQA method, the running time of MSDD is still satisfied and has the potential to be used in real-time applications (about 1.6 seconds for a 720×480 image).

The important message delivered by this paper is that, we demonstrate the feasibility to learn quality-aware features with respect to much smaller-size codebooks while achieving comparable or even better performance. This can be achieved by the following joint efforts: leveraging the quality label of training data as constraints during codebook optimization and making use of the reconstruction residuals in a stage-by-stage manner. The rest of this paper is organized as follows. Section II introduces the related works. Section III illustrates the proposed method with details. Experimental results are presented and analyzed in Section IV. Finally, Section V concludes the paper.

II. Related Work

A. Bag-of-Words (BoW) Model

The BoW model has been widely used for feature representation in image classification [46]. Given a query image, the basic idea of BoW is to quantize each local feature descriptor of this image into the nearest visual word in a codebook, and then represent the image by a histogram of the visual words (i.e., the histogram reflects the distribution of occurred visual words). Finally, the obtained image-level feature vector can be used for classification. Although BoW has undergone significant changes over the past years, it still can be summarized as follows:

Step-1) Codebook generation: First, local feature descriptors are extracted from training images. Next, a visual codebook is learned by seeking a set of representative visual words from the input descriptors. A widely-used method is to perform K-means clustering [50]. Visual words are then defined as the learned clustering centers.

Step-2) Feature encoding: Feature encoding is then performed by embedding local descriptors into the codebook space. This results in so-called feature codes which express each descriptor by a subset of visual words.

Step-3) Spatial pooling: A spatial pooling step involves transforming all the feature codes of an image into a final image-level feature vector called image signature. Finally, training and classification can be performed on the signatures by a discriminative classifier.

Mathematically, the general BoW framework can be formulated as follows

$$\mathbf{x}_l = [x_{l,1},\ x_{l,2}, \cdots,\ x_{l,m}]^T = f(\mathbf{y}_l, \mathcal{D}),\ \ \forall l \in L, \qquad (1)$$

$$\hat{\mathbf{f}} = [f_1, f_2, \cdots,\ f_m]^T,\ \ f_m = g\left(\{x_{l,m}\}_{l \in L}\right), \qquad (2)$$

$$\mathbf{f} = \hat{\mathbf{f}} / \|\hat{\mathbf{f}}\|_2, \qquad (3)$$

where $\mathbf{y}_l \in \mathbb{R}^d$ is an input d-dimensional local feature descriptor to be encoded, $\mathcal{D} = [\mathbf{d}_1,\ \mathbf{d}_2, \cdots, \mathbf{d}_m] \in \mathbb{R}^{d \times m}$ represents the codebook, and m is the codebook size. First, a mapping function $f:\ \mathbb{R}^d \rightarrow \mathbb{R}^m$ in Eq. (1) embeds $\mathbf{y}_l$ into the codebook space resulting in corresponding feature code $\mathbf{x}_l \in \mathbb{R}^m$. Then, spatial pooling operation $g:\ \mathbb{R}^{m \times L} \rightarrow \mathbb{R}^m$ in Eq. (2) aggregates the occurrences of visual words over the entire image: it uses all coefficients $\mathbf{x}_{l,m}$ associated with visual word $\mathbf{d}_m$ to obtain the m-th coefficient in vector $\hat{\mathbf{f}} \in \mathbb{R}^m$. Finally, the signature vector $\hat{\mathbf{f}}$ is normalized in Eq. (3) and the normalized vector representation is fed into a discriminative classifier for classification. Note that the above formulations do not include the codebook generation step as the codebook $\mathcal{D}$ can be constructed by any dictionary learning methods, such as K-means [50] and K-SVD [51].

B. Previous Codebook-based BIQA Methods

The existing codebook-based BIQA methods, including CBIQ [41], CORNIA [42], HOSA [43], and QAF [44]. commonly contain the same components, i.e., local feature extraction, codebook construction, feature encoding, spatial pooling, and quality regression, as depicted in Fig. 1. The codebook construction step is performed off-line. Once a certain codebook is prepared, it will be kept fixed and served as the target feature space over which the feature encoding is implemented. The existing relevant works listed above differ in one or several components with each other. A brief summary of these methods with emphasizing their differences is presented in Table I. In addition, we also add the characteristics of the proposed method here for a better comparison.

The CBIQ method extracts Gabor filter responses from local patches to formulate codebook using K-means and predicts image quality with corresponding codewords occurrence histogram as quality-aware features. Then, SVR is used to learn a quality model that maps such feature vectors to quality scores. However, the size of the built codebook in CBIQ is extremely large, nearly 10K codewords. Later, the same authors extend CBIQ to CORNIA with an unsupervised feature learning method by taking raw image patches as input. With a codebook of length 10K, it obtains considerably performance boost against CBIQ. But when only hundreds of codewords are contained, the performance of CORNIA deteriorates severely. To reduce the codebook size with performance stability, they further design a supervised filter learning (SFL) approach [52] with stochastic gradient descent to optimize a 100-codeword codebook. The performance is acceptable but still inferior to CORNIA. With similar consideration, the quality-aware filter (QAF) method extracts a set of local descriptors from patches to learn a 10K-codeword QAF dictionary using sparse filtering. Another difference between QAF and other codebook-based BIQA models is that random forest (RF) algorithm [53] is used rather than SVR for quality regression. More recently, Xu et al. proposed a novel BIQA method called high order statistic aggregation (HOSA). In HOSA, codebook is also learned by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2763321, IEEE Transactions on Multimedia

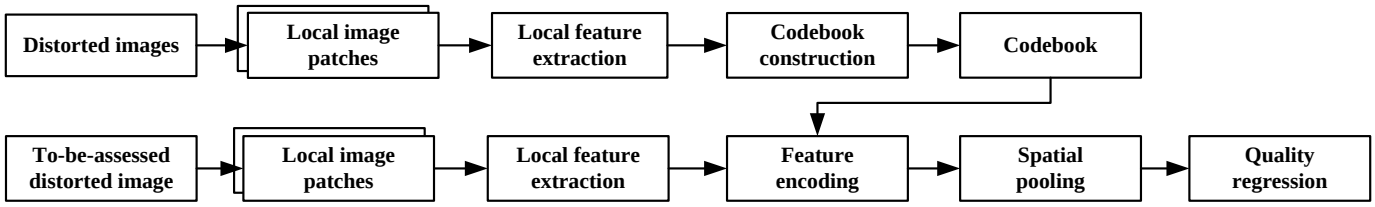


Fig. 1. General framework of the existing codebook-based BIQA methods. The framework mainly contains five components: local feature extraction, codebook construction, feature encoding, spatial pooling, and quality regression. Note that the the codebook construction step is performed offline.

TABLE I
STATE-OF-THE-ART CODEBOOK-BASED BIQA METHODS

Method	Local feature extraction	Codebook construction	Feature encoding	Spatial pooling	Regression
CBIQ [41]	Gabor filter responses	K-means clustering	Hard assignment	Average pooling	SVR
CORNIA [42]	Contrast normalized patches	K-means clustering	Soft assignment	Max pooling	SVR
HOSA [43]	Contrast normalized patches	K-means clustering	High order statistics	Power normalization	SVR
QAF [44]	MSCN and Gabor coefficients	Filter dictionary learning	Sparse filtering	Max pooling	RF
MSDD (Pro.)	Contrast normalized patches	Multi-stage discriminative dictionary learning	Multi-stage feature encoding	Average pooling	SVR

K-means clustering. The main difference between HOSA and the previous relevant ones lies in that, in addition to the mean of each cluster, the dimension-wise variance and skew of clusters are also calculated to form a more comprehensive codebook for feature encoding. With soft-weighted high order statistics difference computation, the final feature dimension in HOSA still reaches to 14.7K.

Overall, the previous codebook-based BIQA methods always extract extremely high-dimensional features for quality prediction, while such high-dimensional features are memory-consuming and have the risk of over-fitting because the image numbers in the existing IQA databases are relatively small as compared to the involved feature dimensionalities. Additionally, all the codebooks in the previous codebook-based BIQA methods are constructed by unsupervised learning algorithms (typically K-means clustering) where the quality information of training samples is not utilized as constraints during codebook optimization. That is, there actually exists a semantic gap between the constructed codebook and BIQA. With such codebook, it is difficult to extract highly discriminative features when the codebook size becomes small because the associated feature discriminability also decreases in this case. This intuitively motivates us to incorporate the quality information of training samples as supervised information into the codebook construction process to optimize a more discriminative codebook for BIQA.

III. PROPOSED METHOD

The diagram of our proposed method is shown in Fig. 2. Overall, the method involves two phases: 1) off-line MSDD learning (MSDDL) and 2) online quality prediction based on MSFE. The first phase, i.e., MSDDL, is implemented off-line and will end when the MSDDs are available. That is, given a testing image, only the second phase is involved. For a certain testing image, after patch partition and contrast normalization processes, the corresponding quality-aware features of each patch can be obtained by MSFE with respect to the learned MSDDs. Then, quality-aware feature codes of each patch are aggregated over the entire image for quality regression using SVR. In essence, the main difference between our proposed method and the previously relevant ones is that that we propose to perform MSFE with respect to multiple cascade and relatively small-size codebooks called MSDDs instead of the traditional single large-size codebook to extract more compact yet discriminative quality-aware features for BIQA.

A. Multi-Stage Discriminative Dictionary Learning (MSDDL)

As revisited, previous codebook-based BIQA methods extract quality-aware features with respect to a single large-size codebook which are typically constructed by unsupervised learning algorithms. In our method, inspired by the motivations of supervised dictionary learning, we propose to incorporate the quality information of local image patches into the traditional dictionary learning frameworks to optimize discriminative dictionary for more compact and effective feature encoding. In addition, this discriminative dictionary learning process is repeated stage-by-stage by further utilizing the reconstruction residual data in each stage until the final stage is reached. We refer to this process as MSDDL hereinafter. Note that the MSDDL is performed off-line. Once the MSDDs are learned, they are kept fixed and served as the target feature space for feature encoding. Given that the optimization of MSDDs requires a set of local patches as well as their corresponding local quality scores as input, an image gallery containing pristine images and their associated distorted images is collected in advance (will be introduced in Section IV-B). Note that the pristine images are also required for local quality estimation using FR-IQA metrics.

1) Contrast Normalization: For each image in the collected image gallery, we convert it into grayscale from which a set of non-overlapping $B_s \times B_s$ image patches $\mathbf{p}_i$ having rich structures and details are sampled. For each patch $\mathbf{p}_i$, the following contrast normalization is performed

$$\mathbf{y}_i = \frac{\mathbf{lum}_i - \mu_i}{\sigma_i + 10}, \tag{4}$$

where $\mathbf{y}_i \in \mathbb{R}^d$ $(d = B_s \times B_s)$, μ_i, and σ_i respective are the contrast normalized patch vector, the mean and the standard

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2763321, IEEE Transactions on Multimedia

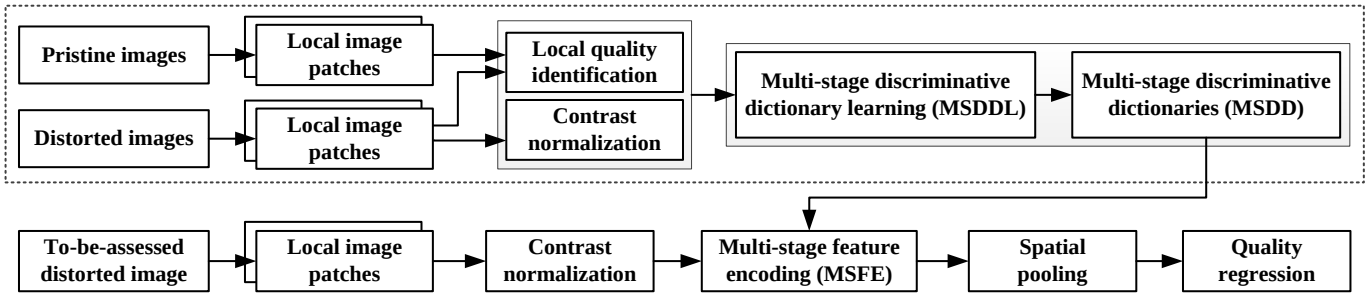


Fig. 2. Framework of the proposed MSDD-based BIQA method. Compared to the previous codebook-based BIQA methods, the main differences of our proposed method lie in codebook construction and feature encoding steps. Specifically, 1) for codebook construction, MSDDL is proposed to replace the traditional unsupervised learning algorithms (e.g., K-means [50] and K-SVD [51]); 2) for feature encoding, MSFE is performed to extract multi-stage feature codes. The feature codes are concatenated across all stages and aggregated over the entire image for quality regression.

deviation of $\mathbf{p}_i$. Note that $\mathbf{lum}_i \in \mathbb{R}^d$ is a vector of intensity values in $\mathbf{p}_i$. In addition, we also perform ZCA whitening to further remove the correlations among local features [54].

2) Local Quality Identification: This step focuses on estimating the quality of each local patch sampled from the image gallery. Recently, many advanced FR quality metrics have shown high consistency with subjective perception, and more importantly, these FR-IQA metrics enable generating pixel-unit quality map. By this consideration, we use the FSIM metric [13] to measure the quality of each local patch. Formally, the FSIM score of $\mathbf{p}_i$ is computed by

$$s_i = \frac{1}{B_s \times B_s} \sum\nolimits_{(x,y)\in \mathbf{p}_i} M_{FSIM_c}(x,y), \tag{5}$$

where $M_{FSIM_c}(x,y)$ represents the FSIM score of pixel (x,y) in $\mathbf{p}_i$ (with a size of $B_s \times B_s$), and $s_i \in [0,1]$ represents the patch-level quality score of $\mathbf{p}_i$, where a higher value of s_i indicates a better quality. In Eq. (5), the pixel-unit FSIM score $M_{FSIM_c}(x,y)$ is computed as follows [55]

$$M_{FSIM_c}(x,y) = \frac{S_{PC}(x,y)\cdot S_G(x,y)\cdot (S_C(x,y))^{\lambda}\cdot PC_{\Delta}(x,y)}{(1/H\times W)\cdot \sum_{x=1}^{H}\sum_{y=1}^{W} PC_{\Delta}(x,y)}, \tag{6}$$

where $S_{PC}(x,y)$, $S_G(x,y)$, and $S_C(x,y)$ respectively denote the quality scores of pixel (x,y) in terms of phase congruency similarity, gradient magnitude similarity, and color information similarity, between the distorted (d) and pristine (o) versions. $PC_{\Delta}(x,y) = max(PC_o(x,y), PC_d(x,y))$ is a pixel-unit weighting map for spatial pooling. PC_o and PC_d are the distorted (d) and pristine (o) phase congruency maps. H and W are the height and width of images. λ is a constant used to adjust the importance of the chromatic components and is set to 0.03 in the default implementation of FSIM in [13]. For brevity, we omit the detailed formulation of $S_{PC}(x,y)$, $S_G(x,y)$, and $S_C(x,y)$.

3) Proposed MSDDL Algorithm: Codebook construction plays a vital role in codebook-based BIQA methods. Generally, the performance accuracy largely depends on the discriminative property of the derived feature codes [56]. It is always true that a better performance will be achieved if providing a much more discriminative codebook for feature encoding. In the literature, many dictionary learning algorithms in unsupervised or supervised manners have been proposed. The first category, unsupervised dictionary learning methods, such as K-means clustering [50] and K-SVD [51], are designed to optimize a codebook for reconstruction purpose instead of classification purpose. That is, codebooks learned via unsupervised learning are not enforced to be discriminative and thus may be suboptimal for classification tasks. By contrast, the second category, i.e., supervised dictionary learning [47], [57], [58], which additionally utilizes the class information of training samples as constraints during dictionary optimization, can offer better solutions to learn dictionaries that are both reconstructive and discriminative and therefore improve the classification accuracy. In view of the competitive performance of the LC-KSVD algorithm [47] in a wide range of classification problems, we utilize LC-KSVD as the basic algorithm unit to design our proposed MSDDL method.

TABLE II
IMPORTANT NOTATIONS AND DEFINITIONS

Notations	Definitions
$\mathbf{P}$	the local image patch set
$\mathbf{Y}$	the contrast normalized patch matrix
$\mathbf{Q}$	the quality-discriminative code matrix
$\hat{\mathbf{D}}$	the learned dictionary
$\hat{\mathbf{X}}$	the estimated sparse code
$\hat{\mathbf{A}}$	the learned linear transformation matrix
$\mathbf{E}$	the reconstruction residual matrix
$\mathbf{c}_k$	the k-th subset
K	the number of subsets
T	the sparsity level
M	the dictionary size
N	the number of stages

Before illustrating the MSDDL algorithm, a summary of some involved notations and definitions are first listed in Table II. Considering all the local patches $\mathbf{P} = \{\mathbf{p}_i\}$ extracted from the image gallery, LC-KSVD is performed by taking $\mathbf{Y} = \{\mathbf{y}_i\}$ and $\mathbf{Q} = \{\mathbf{q}_i\}$ as input, where $\mathbf{y}_i \in \mathbb{R}^d$ and $\mathbf{q}_i \in \mathbb{R}^M$ represent the contrast normalized patch vector and quality-discriminative code of $\mathbf{p}_i$, respectively. Obviously, how to generate the quality-discriminative codes $\mathbf{Q} = \{\mathbf{q}_i\}$ is crucial. First, $\mathbf{Y} = \{\mathbf{y}_i\}$ is grouped into K subsets by

$$\mathbf{c}_k = \left\{ \mathbf{y}_i \middle| \frac{k-1}{K} < s_i \le \frac{k}{K}, \quad k = 1, 2, \cdots, K \right\}, \tag{7}$$

where $\mathbf{c}_k$ denotes the k-th $(k = 1, 2, \cdots, K)$ grouped subset of $\mathbf{Y}$. According to Eq. (7), each $\mathbf{y}_i$ is enforced to be

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2763321, IEEE Transactions on Multimedia

associated with one specific subset $\mathbf{c}_k$. Note that s_i is the estimated quality score of $\mathbf{p}_i$. Then, for each $\mathbf{y}_i$ that is grouped into $\mathbf{c}_k$, its corresponding quality-discriminative code $\mathbf{q}_i$ is determined as follows

$$\mathbf{q}_i = \left[\underbrace{0,0,\cdots,0}_{(M/K)\to\mathbf{c}_1},\cdots,\underbrace{1,1,\cdots,1}_{(M/K)\to\mathbf{c}_k},\cdots,\underbrace{0,0,\cdots,0}_{(M/K)\to\mathbf{c}_K}\right] \in \mathbb{R}^M \tag{8}$$

That is, for each $\mathbf{y}_i$, its corresponding quality-discriminative code $\mathbf{q}_i$ only has (M/K) non-zero entities, where M is the dictionary size and K is the number of grouped subsets. We set $K = 10$ in our implementation.

By taking $\mathbf{Y}$ and $\mathbf{Q}$ as input, the optimization of LC-KSVD can be expressed as

$$\begin{aligned} &\left\langle \hat{\mathbf{D}}, \hat{\mathbf{A}}, \hat{\mathbf{X}} \right\rangle = \arg\min_{\mathbf{D},\mathbf{A},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \alpha \|\mathbf{Q} - \mathbf{A}\mathbf{X}\|_F^2, \\ &s.\,t. \quad \forall i, \quad \|\mathbf{x}_i\|_0 \le T, \end{aligned} \tag{9}$$

where α is defined to control the relative contribution between reconstruction error and quality-discriminative regularization terms. $\hat{\mathbf{D}} = [\mathbf{d}_1,\ \mathbf{d}_2, \cdots, \mathbf{d}_M] \in \mathbb{R}^{d\times M}$ is the learned dictionary (with M atoms) over which $\mathbf{Y} = [\mathbf{y}_1,\ \mathbf{y}_2, \cdots, \mathbf{y}_{N_p}] \in \mathbb{R}^{d\times N_p}$ can have sparse codes $\mathbf{X} = [\mathbf{x}_1,\ \mathbf{x}_2, \cdots, \mathbf{x}_{N_p}] \in \mathbb{R}^{M\times N_p}$, $\hat{\mathbf{A}} = [\mathbf{a}_1,\ \mathbf{a}_2, \cdots, \mathbf{a}_M] \in \mathbb{R}^{M\times M}$ is a linear transformation matrix which transforms the original sparse codes $\mathbf{X}$ to be most discriminative in the sparse feature space $\mathbb{R}^M$, T is a constant specifying the sparsity level, $\|\mathbf{x}_i\|_0$ counts the non-zeros elements in $\mathbf{x}_i$. Owning to the joint consideration of reconstruction error and quality-discriminative regularization, a dictionary that is both reconstructive and discriminative can be learned by optimizing Eq. (9).

As for the purpose of optimization, the objective function defined in Eq. (9) can be rewritten as

$$\begin{aligned} &\left\langle \hat{\mathbf{D}}, \hat{\mathbf{A}}, \hat{\mathbf{X}} \right\rangle = \arg\min_{\mathbf{D},\mathbf{A},\mathbf{X}} \left\| \begin{pmatrix} \mathbf{Y} \\ \sqrt{\alpha}\mathbf{Q} \end{pmatrix} - \begin{pmatrix} \mathbf{D} \\ \sqrt{\alpha}\mathbf{A} \end{pmatrix} \mathbf{X} \right\|_F^2, \\ &s.\,t. \quad \forall i, \quad \|\mathbf{x}_i\|_0 \le T, \end{aligned} \tag{10}$$

Let $\mathbf{Y}_{new} = (\mathbf{Y}^T, \sqrt{\alpha}\mathbf{Q}^T)^T$, $\mathbf{D}_{new} = (\mathbf{D}^T, \sqrt{\alpha}\mathbf{A}^T)^T$. Note that $\mathbf{D}_{new}$ is ℓ_2 normalized column-wise. Thus, the optimization of Eq. (10) is equivalent to solve the following problem

$$\begin{aligned} &\left\langle \hat{\mathbf{D}}_{new}, \hat{\mathbf{X}} \right\rangle = \arg\min_{\mathbf{D}_{new},\mathbf{X}} \| \mathbf{Y}_{new} - \mathbf{D}_{new}\mathbf{X} \|_F^2, \\ &s.\,t. \quad \forall i, \quad \|\mathbf{x}_i\|_0 \le T, \end{aligned} \tag{11}$$

which is exactly the problem that standard K-SVD [51] solves. To solve this problem with K-SVD, both $\mathbf{D}$ and $\mathbf{A}$ need to be initialized as $\mathbf{D}^0$ and $\mathbf{A}^0$, respectively. Due to the space limit, we omit the initialization process here. The details can be found in [47]. Finally, we get $\hat{\mathbf{D}}_{new}$ and $\hat{\mathbf{X}}$ by solving Eq. (11) using K-SVD. From $\hat{\mathbf{D}}_{new}$, we can further obtain $\hat{\mathbf{D}} = [\mathbf{d}_1,\ \mathbf{d}_2,\ \ldots,\ \mathbf{d}_M]$ and $\hat{\mathbf{A}} = [\mathbf{a}_1,\ \mathbf{a}_2, \cdots, \mathbf{a}_M]$. However, they cannot be directly used for testing because $\hat{\mathbf{D}}$ and $\hat{\mathbf{A}}$ are jointly normalized in $\hat{\mathbf{D}}_{new}$ during LC-KSVD optimization, i.e., $\forall m,\ ||\mathbf{d}_m^T, \sqrt{\alpha}\mathbf{a}_m^T||_2^2 = 1$. The desired dictionaries $\hat{\mathbf{D}}^*$ and $\hat{\mathbf{A}}^*$ can be computed as

$$\hat{\mathbf{D}}^* = \left[\frac{\mathbf{d}_1}{\|\mathbf{d}_1\|_2}, \frac{\mathbf{d}_2}{\|\mathbf{d}_2\|_2}, \cdots, \frac{\mathbf{d}_M}{\|\mathbf{d}_M\|_2}\right], \tag{12}$$

$$\hat{\mathbf{A}}^* = \left[\frac{\mathbf{a}_1}{\|\mathbf{a}_1\|_2}, \frac{\mathbf{a}_2}{\|\mathbf{a}_2\|_2}, \cdots, \frac{\mathbf{a}_M}{\|\mathbf{a}_M\|_2}\right]. \tag{13}$$

Based on the desired dictionary $\hat{\mathbf{D}}^*$ and the sparse codes $\mathbf{X}$, the residual matrix $\mathbf{E}$ of reconstructing $\mathbf{Y}$ is given by

$$\mathbf{E} = \mathbf{Y} - \hat{\mathbf{D}}^*\mathbf{X}. \tag{14}$$

To make use of the reconstruction residuals, our proposed MSDDL method applies LC-KSVD repeatedly to optimize N-stage discriminative dictionaries $\{\hat{\mathbf{D}}_1^*, \hat{\mathbf{D}}_2^*, \cdots, \hat{\mathbf{D}}_N^*\}$. The optimization of LC-KSVD in the $(n+1)$-th stage begins only when the n-th stage ends. Once the optimization of LC-KSVD in the n-th stage is finished, the generated dictionary $\hat{\mathbf{D}}_n^*$ is stored while the reconstruction residual matrix $\mathbf{E}_n$ is passed into the $(n+1)$-th stage as inputs based on which a new round of LC-KSVD optimization is performed. This kind of iteration stops until the final stage is reached. To facilitate understanding, we summarize the pseudo-code of our proposed MSDDL method in Algorithm 1.

Algorithm 1 Multi-stage discriminative dictionary learning.

Input: $\mathbf{Y}$; $\mathbf{Q}$; α; T; M; N;
Output: $\hat{\mathbf{D}}^* = \{\hat{\mathbf{D}}_1^*,\ \hat{\mathbf{D}}_2^*,\ \cdots,\ \hat{\mathbf{D}}_N^*\}$;
1: initialize $\mathbf{Y}_1 = \mathbf{Y}$;
2: **for** each $n \in [1, N]$ **do**
3: do LC-KSVD [47] on $\mathbf{Y}_n$ according to Eq. (9);
4: output the n-th stage dictionary $\hat{\mathbf{D}}_n^*$ and the sparse code $\mathbf{X}_n$;
5: compute the residual matrix $\mathbf{E}_n$ according to Eq. (14);
6: pass $\mathbf{E}_n$ to the next stage as input: $\mathbf{Y}_{n+1} = \mathbf{E}_n$;
7: **end for**
8: concatenate the dictionaries in all stages: $\hat{\mathbf{D}}^* = \{\hat{\mathbf{D}}_1^*,\ \hat{\mathbf{D}}_2^*,\ \cdots,\ \hat{\mathbf{D}}_N^*\}$.

B. Multi-Stage Feature Encoding (MSFE)-Based BIQA

Feature encoding can be understood as an activation function for the learned dictionary. By feature encoding, new feature representations can be obtained with the transformation from original feature space to the target dictionary space; the associated activities of atoms are the resultant feature codes. Given the learned MSDDs, feature encoding is also performed in a stage-by-stage manner. Therefore, the feature encoding process is referred to as MSFE in this paper. For an arbitrary local patch $\mathbf{p}$ extracted from the testing image, we first compute its contrast normalized patch vector $\mathbf{y}$ according to Eq. (4). By taking $\mathbf{y}$ as input, the MSFE operates as follows. On the one hand, the sparse codes $\mathbf{x}$ of feature $\mathbf{y}$ over the pre-learned MSDDs is estimated in each stage. On the other hand, reconstruction residual vector in the current stage is computed and passed to the next stage. Given a set of N-stage dictionaries $\{\hat{\mathbf{D}}_1^*,\ \hat{\mathbf{D}}_2^*,\ \cdots,\ \hat{\mathbf{D}}_N^*\}$, for each $\mathbf{y}_n \in \mathbb{R}^d$ as inputs in the n-th stage, its corresponding sparse code $\mathbf{x}_n \in \mathbb{R}^M$ over $\hat{\mathbf{D}}_n^*$ and reconstruction residual $\mathbf{e}_n \in \mathbb{R}^d$ are respectively calculated as follows

$$\mathbf{x}_n = \arg\min_{\mathbf{x}_n} \left\|\mathbf{y}_n - \hat{\mathbf{D}}_n^*\mathbf{x}_n\right\|_F^2, \quad s.\,t. \quad \|\mathbf{x}_n\|_0 \le T, \tag{15}$$

Algorithm 2 Multi-stage feature encoding.

Input: $\hat{\mathbf{D}}^* = \{\hat{\mathbf{D}}_1^*,\ \hat{\mathbf{D}}_2^*,\ \cdots,\ \hat{\mathbf{D}}_N^*\}$; $\mathbf{y}$;
Output: $\mathbf{x} = [\mathbf{x}_1,\ \mathbf{x}_2,\ \cdots,\ \mathbf{x}_N]$;
1: initialize $\mathbf{y}_1 = \mathbf{y}$;
2: **for** each $n \in [1, N]$ **do**
3: do sparse coding [59] on $\mathbf{y}_n$ according to Eq. (15);
4: output the sparse code $\mathbf{x}_n$;
5: compute the residual vector $\mathbf{e}_n$ according to Eq. (16);
6: pass $\mathbf{e}_n$ to the next stage as input: $\mathbf{y}_{n+1} = \mathbf{e}_n$;
7: **end for**
8: concatenate the sparse codes in all stages: $\mathbf{x} = [\mathbf{x}_1,\ \mathbf{x}_2,\ \cdots,\ \mathbf{x}_N]$.

$$\mathbf{e}_n = \mathbf{y}_n - \hat{\mathbf{D}}_n^* \cdot \mathbf{x}_n, \tag{16}$$

To solve Eq. (15), we resort to the batch-OMP algorithm [59] to obtain the optimal solution. Once the sparse code $\mathbf{x}_n$ and the reconstruction residual $\mathbf{e}_n$ are estimated at the current stage, $\mathbf{e}_n$ is passed to the next stage, i.e., $\mathbf{y}_{n+1} = \mathbf{e}_n$. Afterwards, the steps in Eq. (15) and (16) are iteratively performed until the final stage is reached. Finally, the obtained sparse code vectors $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ in all stages are concatenated, yielding the final feature code to represent $\mathbf{y}$, i.e., $\mathbf{y} \in \mathbb{R}^d \rightarrow \mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]^T \in \mathbb{R}^{(M \times N) \times 1}$, where N is the number of stages and M is the number of atoms in each single-stage dictionary. The pseudo-code of MSFE is summarized in Algorithm 2.

It is worthy emphasizing that the final feature code comes from multi-stage sparse code vectors over MSDDs which are enforced to be both reconstructive and discriminative with the help of multi-stage LC-KSVD optimization. On the one hand, such kind of dictionary could enable better discriminability of the estimated sparse codes for quality prediction. On the other hand, the multi-stage framework could make use of the reconstruction residuals which also provide complementary benefits for characterizing image quality.

C. Spatial Pooling

The feature codes of all the local patches extracted from a testing image should be aggregated to get a final image-level feature vector convenient for quality regression. We resort to the simple average-pooling for the sake of high efficiency. The benefits of other advanced pooling strategies can be further exploited [60]. To be more specific, given a testing image, we similarly partition it into non-overlapped patches $\{\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_L\}$ and compute the corresponding contrast normalized patch vectors $\{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_L\}$ according to Eq. (4). Then, following the MSFE process described in Algorithm 2, the associated patch-level feature codes of this image are obtained, as denoted by $\{\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^L\}$ where $\mathbf{x}^l = [x_1^l, x_2^l, \cdots, x_{M \times N}^l]^T \in \mathbb{R}^{(M \times N) \times 1}, l = 1, 2, \cdots, L$. Finally, the average-pooling is expressed as

$$f_i = \frac{1}{L} \sum\nolimits_{l=1}^{L} x_i^l, \quad i = 1,\ 2,\ \cdots,\ (M \times N), \tag{17}$$

where $\mathbf{f} = [f_1, f_2, \cdots, f_{M \times N}]$ represents the final image-level quality-aware feature vector.

D. Quality Regression

After feature extraction, the quality evaluation is achieved using SVR to create a fair comparison with state-of-the-art BIQA methods. Specifically, a SVR model is first learned using a set of training images. Then the trained SVR model is used to evaluate the quality of testing images. We utilize the LIBSVM package [48] to implement the SVR with radial basis function (RBF) as the kernel.

IV. Experimental Results

A. Evaluation Protocols

The experiments are conducted on five databases: LIVE [61], CSIQ [10], TID2013 [49], LIVEMD [62], and SIQAD [63]. Following protocols in CBIQ [41], CORNIA [42], and HOSA [43], all five distortion types in LIVE (i.e., JP2K, JPEG, WN, GB, and FF) are considered in the experiments, while for CSIQ, TID2013, and SIQAD, only the common four distortion types appeared in LIVE (i.e., JP2K, JPEG, WN and GB) are considered. Given that the proposed method belongs to the codebook-based category, we compare it with three codebook-based BIQA methods, i.e., CBIQ [41], CORNIA [42], and HOSA [43]. In addition, several representative handcrafted NSS eature-based BIQA methods, i.e., DIIVINE [33], BLIINDS-II [34], BRISQUE [35], GM-LOG [37], and NFREM [38], are also included for comparison.

Three commonly-used criteria, i.e., Spearmans rank order correlation coefficient (SROCC) which measures the prediction monotonicity, Pearsons linear correlation coefficient (PLCC) which measures the prediction accuracy, and root mean squared error (RMSE) which measures the prediction error, are used to evaluate the performance. A good BIQA model is excepted to have larger values of SROCC and PLCC, while a smaller value of RMSE. As recommended by the report from Video Quality Expert Group (VQEG) [64], the relationship between the subjective scores and the objective scores may not be linear due to the nonlinear quality rating of observers. Therefore, before calculating PLCC and RMSE, a nonlinear logistic regression process is applied

$$f(x) = \beta_1 \left[\frac{1}{2} - \frac{1}{1 + e^{\beta_2 (x - \beta_3)}}\right] + \beta_4 x + \beta_5, \tag{18}$$

where β_1, β_2, β_3, β_4, and β_5 are the parameters to be fitted.

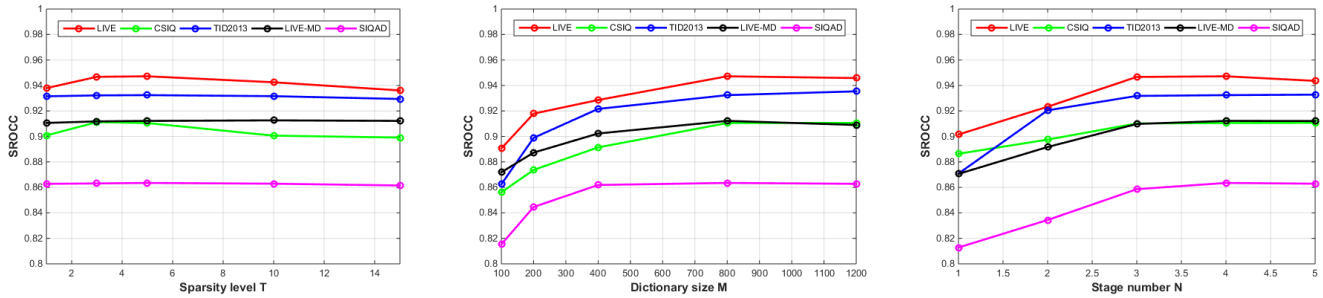
B. Implementation Details

To implement MSDDL, an image gallery set containing pristine images and their associated distorted images should be prepared in advance. We collect 12 pristine images which have different scenes from the images appeared in existing IQA databases, as shown in Fig. 3. Based on these pristine images, the distorted images are generated by simulating four distortion types (i.e., JP2K, JPEG, GB, and WN) with five quality degradation levels. As a result, 240 distorted images with different distortions and qualities are obtained. To learn MSDDs, a total number of $1,000$ non-overlapped patches having rich structures (with the largest variance values) are sampled from each distorted image to form $\mathbf{Y}$. The rational is that, image patches with larger variance values are considered to have richer structure information which is responsible the most for informative dictionary learning. For the selection of FSIM metric in our method, we experimentally tried several

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2763321, IEEE Transactions on Multimedia

(a) Collected pristine images (b) Pristine images in LIVE (c) Pristine images in CSIQ (d) Pristine images in TID2013

Fig. 3. Pristine images. From left to right: Collected pristine images, pristine images in LIVE, CSIQ, and TID2013, respectively.



(a): SROCC vs. T (b): SROCC vs. M (c): SROCC vs. N

Fig. 4. Performance results in terms of SROCC vs. (a) the sparsity level T; (b) dictionary size M; (c) the number of stages N.

leading FR-IQA metrics (i.e., SSIM [9], GMSD [16], FSIM [13], and VSI [15]). Similar to FSIM, VSI also only provide a single quality score for each image after pooling, then the pixel-unit local quality map based on VSI is estimated by a similar formulation defined in Eq. (6) (just replace the similarity measures in Eq. (6) by the relevant ones in VSI). In contrast to FSIM and VSI, SSIM and GMSD are able to provide local quality maps in its original form, then they can be applied to construct the local quality map seamlessly. By experiments, we find that 1) the FSIM metric is suitable in our problem, and 2) the influence of different leading FR-IQA metrics is not obvious.

In our method, several parameters are involved: 1) B_s: patch size; 2) K: number of grouped subsets; 2) T: sparsity level; 3) M: dictionary size; 4) N: number of stages. In our experiments, the patch size B_s is set to 7, which is the same with CORNIA and HOSA. The number of subsets K is selected from a set of candidates $K_\Omega = \{5, 10, 15, 20\}$. Among these candidates, the one associated with the highest SROCC value on LIVE is selected and then used for all the databases. Although this final $K = 10$ may not be optimal for all the databases (the optimal K can be quite different for each database actually), it still leads to promising performance for all cases, as confirmed by all the experimental results. To investigate the influence of sparsity level T, dictionary size M, and number of stages N on the performance, we assign $T = \{1, 3, 5, 10, 15\}$, $M = \{100, 200, 400, 800, 1200\}$, $N = \{1, 2, 3, 4, 5\}$ and compute the SROCC values over all five databases. The results are shown in Fig. 4. For all databases, we set $T = 5$, $M = 800$, $N = 4$, which can achieve satisfactory performances for most cases. Finally, the overall feature dimension of our method is 3.2K in total which is much smaller than 10K used in CBIQ, 20K used in CORNIA and 14.7K used in HOSA. Such feature dimension reduction is meaningful for the applications of embedding systems and mobile devices where the memory resources are limited. It is reminded that, although the parameters are not determined on a separate training set, they are also not optimized for each database. By observing the performance variations to different parameter choices, we simply choose the parameters that generally work well for most cases. As the selected parameters can achieve satisfied performance across all the five databases, it is believed that such parameters can be directly used for other testing samples.

C. Performance on Individual Database

To evaluate the performance of the BIQA algorithms, we test them on each individual database separately. Following the evaluation protocols in previous works, we divide each database into training and testing subsets. To ensure that there are non-overlapping contents used for training and testing, distorted images associated with 80% of the reference images in each database are selected for training, and the rest 20% distorted images are used for testing. Such random training-testing split is repeated $1,000$ times and the median performance is reported in Table III.

From Table III, we have the following observations.

First, the results of all methods are worse (some less and some more) on LIVE MD and SIQAD. Such performance bias can be explained as follows. For LIVE MD, the challenges can be summarized from three aspects: the influence of individual distortions on image quality, the interaction between these distortions, and the joint effects of theses distortions on the overall quality. All these problems make the quality evaluation of images degraded with mixed distortions challenging. For

TABLE III
OVERALL PERFORMANCE OF DIFFERENT BIQA MODELS. TEXTS IN BOLD INDICATE THE TOP THREE METHODS.

Database	Criteria	Handcrafted feature-based methods					Codebook feature-based methods				
		DIIVINE	BLIINDS-II	BRISQUE	GM-LOG	NFREM	CBIQ	CORNIA	CORNIA-1.6K	HOSA	MSDD
LIVE	SROCC	0.9162	0.9302	0.9409	**0.9503**	0.9376	0.9119	0.9417	0.9174	**0.9504**	**0.9472**
	PLCC	0.9172	0.9357	0.9450	**0.9539**	0.9421	0.9278	0.9434	0.9192	**0.9527**	**0.9488**
	RMSE	10.8103	9.6189	8.9048	**8.1723**	9.1012	10.2683	9.0204	10.6945	**8.2858**	**8.7719**
CSIQ	SROCC	0.8760	0.9140	0.9099	**0.9228**	**0.9287**	0.9048	0.8930	0.8817	**0.9298**	0.9105
	PLCC	0.8983	0.9323	0.9278	**0.9408**	**0.9518**	0.8922	0.9175	0.9024	**0.9480**	0.9227
	RMSE	0.1220	0.1010	0.1044	**0.0950**	**0.0842**	0.1225	0.1123	0.1181	**0.0887**	0.1080
TID2013	SROCC	0.8753	0.8786	0.8917	0.9282	**0.9286**	0.8946	0.8998	0.8878	**0.9521**	**0.9324**
	PLCC	0.8859	0.9053	0.9176	**0.9439**	**0.9509**	0.8837	0.9277	0.9175	**0.9592**	0.9416
	RMSE	0.6474	0.5921	0.5534	**0.4629**	**0.4363**	0.6491	0.5239	0.5533	**0.3941**	0.4687
LIVE-MD	SROCC	0.8738	0.8872	0.8972	0.8237	0.8989	0.8876	**0.9007**	0.8916	**0.9019**	**0.9122**
	PLCC	0.8936	0.9028	**0.9207**	0.8632	0.9190	0.8992	0.9150	0.9038	**0.9262**	**0.9325**
	RMSE	8.3843	8.1330	**7.3168**	9.4198	7.4132	8.1455	7.6737	8.1295	**6.9739**	**6.9476**
SIQAD	SROCC	0.7279	0.7561	0.7715	0.7989	0.7983	0.8314	**0.8352**	0.8238	**0.8484**	**0.8635**
	PLCC	0.7768	0.7982	0.8210	0.8330	0.8259	0.8492	**0.8533**	0.8365	**0.8636**	**0.8792**
	RMSE	8.6903	8.3688	7.9383	7.7005	7.8615	7.2571	**7.1989**	7.6912	**6.9594**	**6.6733**

SIQAD, the main challenge lies in that images are of computer graphic or document contents, not resulting from a natural source. Therefore, handcrafted NSS feature-based methods may lose their power in evaluating the quality of screen content images. Given that the influences of the textual contents and the one of the pictorial contents on the overall quality can be quite different, such quality evaluation problem especially a blind one is also challenging.

Second, the proposed MSDD method ranks the top three for most databases (except for the CSIQ) and ranks top for six times in total followed by the HOSA for five times. This demonstrates that MSDD has fairly good ability to learn effective yet robust features in characterizing image quality with diverse visual contents and distortions.

Third, compared to the handcrafted NSS feature-based methods, e.g., BRISQUE, GM-LOG, and NFREM, the codebook-based BIQA methods show comparable performance on LIVE, CSIQ, and TID2013 which only contain singly-distorted natural images, while for LIVEMD and SIQAD, especially SIQAD, the handcrafted NSS feature-based methods are significantly inferior to all the codebook-based ones. This indicates that traditional handcrafted NSS features cannot well reflect the quality information contained in distorted screen content images which contain not only natural pictorial contents but also textual and graphical contents. However, the codebook-based methods can provide unique benefits in capturing the specific information in screen content images.

Forth, compared to the previous codebook-based BIQA methods, i.e., CBIQ, CORNIA, and HOSA, MSDD delivers competitive performance on LIVE, CSIQ, and TID2013, while performs better on LIVE-MD and SIQAD. It is encouraging because MSDD only need to extract a 3.2K-dimensional feature vector to represent an image while previous codebook-based methods extract 10K-dimensional feature vector at least. Although CORNIA-1.6K also relies on 3.2K-dimensional features, it performs much worse than our method across all databases. We believe this phenomenon is mainly attributed to the large encoding error with a single small-size codebook which may inevitably interfere with the accurate evaluation of quality degradation. Instead, the proposed method encodes each local patch with respect to multiple cascade discriminative dictionaries, yielding more compact and discriminative quality-aware features.

D. Statistical Significance Test

To understand whether the advantages of our method over the competing methods are statistically significant, we further evaluate the statistical significance using the Wilcoxon rank-sum test [65] which measures the equivalence of the median values of two independent samples. We conduct the Wilcoxon rank-sum test at a significance level of 5% using the 1000 SROCC values of all pairs of BIQA methods. The null hypothesis of this analysis assumes that the SROCC values of the methods in comparison are drawn from populations with equal means. The tables in Fig. 5 provides the results over five databases. In the tables, a symbol of '1' indicates that the row model is statistically superior to the column model, a symbol of '-1' indicates that the row model is statistically inferior, and a symbol of '0' indicates that the row model is statistically equivalent to the compared model in that column.

From the tables, we can see that the proposed method performs statistically better than all the other competing methods on both LIVE-MD and SIQAD. Furthermore, our method is on par with BRISQUE, GM-LOG, CORNIA, and HOSA, while significantly superior to the rest competing methods on LIVE. As for CSIQ, the proposed method is statistically equivalent to BLIINDS-II, BRISQUE, and CBIQ, however, significantly worse than GM-LOG, NFREM, and HOSA. For TID2013, only HOSA is significantly better than the proposed method. Overall speaking, the proposed method is comparable with HOSA when considering the significance test results over all these databases (symbol '1' occurs 31 times for the proposed method and 36 times for HOSA on all five databases). However, given the fact that much lower dimension features are required in our method as compared to HOSA (3.2K vs. 14.7K), the reported performance is quite competitive in terms of either efficacy or efficiency. Note that, when the same dimension features are used, our method is significantly better on all databases, as demonstrated by the

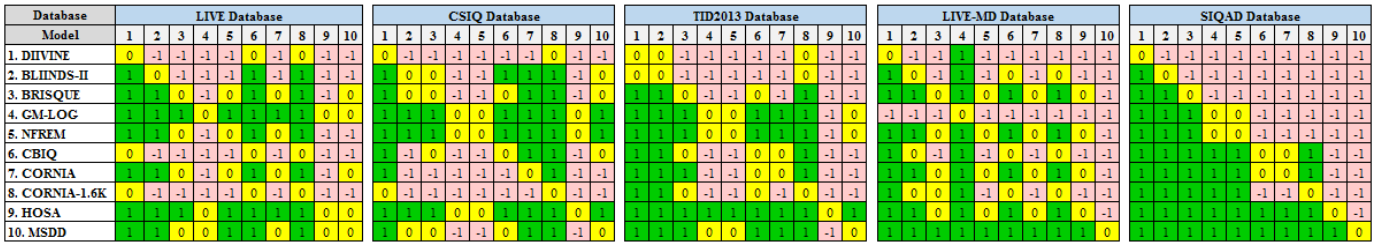


Fig. 5. Statistical significance test results of competing BIQA methods. In the Table, "1" indicates the row model is statistically better than the column model; "-1" indicates the row model is statistically worse than the column model; "0" indicates the row and column models are statistically equivalent.

TABLE IV
PERFORMANCE (SROCC) OF DIFFERENT MODELS ON INDIVIDUAL DISTORTION TYPES. TEXTS IN BOLD INDICATE THE TOP THREE METHODS.

Database	Distortion	DIIVINE	BLIINDS-II	BRISQUE	GM-LOG	NFREM	CBIQ	CORNIA	CORNIA-1.6K	HOSA	MSDD
LIVE	JP2K	0.9164	0.9301	0.9169	0.9262	**0.9368**	0.9033	0.9211	0.8966	**0.9331**	**0.9328**
	JPEG	0.9028	0.9505	**0.9650**	**0.9631**	**0.9641**	0.9418	0.9382	0.9074	0.9549	0.9311
	WN	**0.9813**	0.9471	0.9800	**0.9831**	**0.9838**	0.9321	0.9568	0.9362	0.9729	0.9572
	GB	0.9299	0.9146	**0.9519**	0.9293	0.9091	0.9345	**0.9573**	0.9408	**0.9524**	0.9485
	FF	0.8627	0.8741	0.8754	**0.8994**	0.8502	0.8558	**0.9057**	0.8843	**0.9003**	0.8922
CSIQ	JP2K	0.8662	0.9052	0.8952	**0.9177**	**0.9167**	0.9008	0.9055	0.8819	**0.9244**	0.9033
	JPEG	0.8802	**0.9254**	0.9248	0.9161	**0.9266**	0.9063	0.8888	0.8672	**0.9254**	0.9149
	WN	0.9034	**0.9368**	**0.9379**	**0.9471**	0.9338	0.9209	0.8080	0.7848	0.9192	0.8857
	GB	0.8754	0.9164	0.9123	0.9132	**0.9240**	0.8742	0.9066	0.8993	**0.9266**	**0.9237**
TID2013	JP2K	0.8662	0.9016	0.9011	0.9280	**0.9375**	0.8716	0.9123	0.8945	**0.9453**	**0.9291**
	JPEG	0.8685	0.8546	0.8723	**0.9084**	0.8959	0.9050	0.8654	0.8471	**0.9283**	**0.9052**
	WN	0.8845	0.8315	0.8568	**0.9385**	**0.9391**	0.8948	0.7546	0.7368	**0.9215**	0.9166
	GB	**0.9369**	0.8731	0.9201	0.9192	0.9282	0.9022	0.9234	0.9115	**0.9538**	**0.9377**
LIVE-MD	GB+JPEG	0.8773	0.8993	0.9029	0.8237	0.9188	0.8914	**0.9006**	0.8836	**0.9287**	**0.9218**
	GB+WN	0.8819	0.8898	**0.9022**	0.8632	0.8874	0.8837	**0.8991**	0.8852	0.8918	**0.9011**
SIQAD	JP2K	0.4527	0.6234	0.4466	0.6716	0.6635	0.7236	**0.7348**	0.7259	**0.7701**	**0.7945**
	JPEG	0.3519	0.3755	0.5690	0.4442	0.4369	0.7412	**0.7682**	**0.7527**	0.7523	**0.7864**
	WN	0.8528	**0.8708**	0.8621	**0.8889**	0.8703	0.8353	0.8404	0.8384	0.8530	**0.8812**
	GB	**0.8990**	0.8626	**0.8963**	0.8768	0.8758	0.8691	0.8736	0.8675	0.8840	**0.8941**
Hit count		3	3	5	8	9	0	5	1	12	11

significance test results between the proposed method and CORNIA-1.6K (symbol '1' only occurs 9 times for CORNIA-1.6K over all five databases).

E. Performance on Individual Distortion Type

Besides the overall performances on entire databases, we are also interested to know the performances on individual distortion types. For each individual distortion type, we test the images that belong to each distortion type in the testing set with the model trained on 80% of images including all types of distortions in that database. The results are presented in Table IV and the best three results are highlighted in boldface. For brevity, we only report SROCC results without the loss of generality. In addition, we show the hit count (i.e., the number of times ranked in the top three for each distortion type) of the performance for each competing method. It is seen that HOSA has the highest hit count (12 times), followed by the proposed method (11 times) and NFREM (9 times). Although comparable, our method depends on much lower dimensional features than HOSA and performs more stable than NFREM and GM-LOG over all these databases especially on LIVE-MD and SIQAD databases.

To further validate the general capacity, we also conducted the 1000 train-test experiments for all competing BIQA models on the entire TID2013 database. The distortion types in TID2013 database include: #01 additive white Gaussian noise, #02 additive noise in color components, #03 additive Gaussian spatially correlated noise, #04 masked noise, #05 high-frequency noise, #06 impulse noise, #07 quantization noise, #08 Gaussian blur, #09 image denoising, #10 JPEG compression, #11 JPEG2000 compression, #12 JPEG transmission errors, #13 JPEG2000 transmission errors, #14 non eccentricity pattern noise, #15 local block-wise distortion of different intensity, #16 mean shift, #17 contrast change, #18 change of color saturation, #19 multiplicative Gaussian noise, #20 comfort noise, #21 lossy compression of noisy images, #22 image color quantization with dither, #23 chromatic aberrations and #24 sparse sampling and reconstruction. For individual distortion type, we tested the images belonging to each distortion in the testing set with the model trained on 80% of images including all types of distortions in the entire TID2013 database. The results are summarized in Table V and the best three results are highlighted in boldface. For brevity, we only present the SROCC results. Similar conclusions can be obtained for PLCC and RMSE.

From Table V, the following observations can be found. First, MSDD performs continuously better than both CORNIA and CORNIA-1.6K which further validates its effectiveness. Second, MSDD is slightly worse than HOSA on the entire TID2013 database, while for each individual distortion type, MSDD performs worse than HOSA 16 times and better 7 times. Nevertheless, when considering the fact that much less

TABLE V
PERFORMANCE EVALUATION (SROCC) ON THE ENTIRE TID2013 DATABASE. EACH NUMBER CORRESPONDS TO A SPECIFIC DISTORTION TYPE.

Method	#01	#02	#03	#04	#05	#06	#07	#08	#09	#10	#11	#12	#13
DIIVINE	0.5831	0.3240	0.6005	0.3214	0.7649	0.6225	0.5986	0.8337	0.7229	0.6973	0.8228	0.4316	0.5232
BLIINDS-II	0.7142	**0.7282**	**0.8245**	0.3577	0.8523	0.6641	0.7799	0.8523	0.7538	**0.8077**	0.8615	0.2512	**0.7550**
BRISQUE	0.6300	0.4235	0.7265	0.3210	0.7754	0.6692	0.5915	0.8446	0.5533	0.7417	0.7988	0.3012	0.6715
GM-LOG	**0.7808**	**0.5881**	**0.8177**	**0.5449**	**0.8892**	0.6593	**0.8000**	0.8485	0.7531	0.7992	0.8431	0.3985	**0.7473**
NFREM	**0.8508**	0.5201	**0.8458**	**0.5209**	**0.8936**	**0.8573**	0.7848	0.8876	0.7412	0.7972	**0.9195**	0.3807	**0.7181**
CORNIA	0.3408	-0.1962	0.6892	0.1835	0.6071	-0.0138	0.6731	**0.8957**	**0.7866**	0.7854	0.8831	**0.5515**	0.5469
CORNIA-1.6K	0.5591	0.1343	0.5372	0.2269	0.6651	0.1877	0.6385	0.8716	0.7695	0.6872	0.8527	0.4683	0.5069
HOSA	**0.8529**	**0.6250**	0.7820	0.3677	**0.9046**	**0.7746**	**0.8101**	**0.8924**	**0.8702**	**0.8931**	**0.9323**	**0.7472**	0.7012
MSDD (Pro.)	0.6519	0.4870	0.7885	**0.3718**	0.7772	**0.6855**	**0.8023**	**0.9022**	**0.8236**	**0.8446**	**0.9207**	**0.6059**	0.6431
Method	#14	#15	#16	#17	#18	#19	#20	#21	#22	#23	#24	All	Hit count
DIIVINE	**0.3114**	0.1998	**0.2754**	0.0315	**0.2008**	0.6012	0.2102	0.4887	0.6332	0.7616	0.8389	0.5296	3
BLIINDS-II	0.0812	**0.3713**	0.1585	-0.0823	0.1092	0.6987	0.2223	0.4505	0.8146	0.5676	0.8562	0.5504	5
BRISQUE	0.1751	0.1835	0.1545	0.1246	0.0315	0.5596	0.2823	**0.6803**	0.8038	0.7145	0.7995	0.5615	1
GM-LOG	**0.2054**	**0.2419**	0.0758	0.2946	-0.1831	**0.7246**	0.2502	**0.6419**	**0.8565**	0.6582	**0.9031**	**0.6750**	13
NFREM	0.1758	0.0812	**0.2380**	0.0559	-0.0287	**0.7621**	0.2064	0.4007	**0.8482**	0.6838	0.8781	0.6522	10
CORNIA	0.1605	0.0962	0.0077	**0.4233**	-0.0554	0.2593	**0.6064**	0.5546	0.5919	**0.7592**	0.9023	0.6509	6
CORNIA-1.6K	0.1721	0.0815	0.0248	**0.4056**	0.0612	0.4281	**0.5482**	0.5966	0.5875	0.7321	0.8819	0.5867	2
HOSA	0.1989	**0.3273**	**0.2327**	0.2938	**0.1185**	**0.7819**	0.5315	**0.8354**	**0.8554**	**0.8014**	**0.9052**	**0.7280**	18
MSDD (Pro.)	**0.2074**	0.1452	0.2086	**0.4235**	**0.1209**	0.3864	**0.6237**	0.6018	0.6775	**0.7810**	**0.9045**	**0.7033**	15

features are involved in MSDD as compared to HOSA, such performance is also encouraging. It is reminded that the feature dimensionality in HOSA is difficult to be reduced as even a 100-codeword codebook is used, a total number of 14,700-dim feature vector will be produced. Third, most methods deliver relatively satisfied performance on evaluating the noise-related, such as #1, #3, #5, #6, #7, #9, and compression-related distortions, such as #10, #11, #24, while failing to evaluate several color- and contrast-related distortions, such as #2, #14, #15, #16, #17, #18, #22, and #23 in TID2013. This is very challenging due to these reasons: 1) #2, #18, #22, and #23 is mainly about color saturation thus most BIQA methods which based on luminance image processing fail to accurately estimate the resultant quality; 2) #14 and #15 consist of localized distortion patterns which have limited influence on global image feature; 3) #16 and #17 are correlated to image luminance change which is generally overlooked since current algorithms always work on normalized images. All these failed cases could lead us to develop more robust and universal BIQA models in the future by considering more comprehensive quality-aware features.

F. Dependency on Training Set Size

Each database is divided into two non-overlapping subsets for performance evaluation, i.e., 80% samples for training and the remaining 20% samples for testing. To investigate the dependency of model performance to different training set sizes, we measure the mean SROCC and PLCC values over 1000 random train-test splits as a function of the percentage of training set sizes, ranging from 10% to 90% with an increment of 10%. The results are shown in Fig. 6. It is notable that the proposed method does not deteriorate substantially along with the reduction of the training set size. Specifically, the performance reduction in terms of SROCC (PLCC) caused by decreasing the percentage of training set from 90% to 30% are less than 0.03 (0.04) for all five databases, indicating the robustness against different training set sizes.

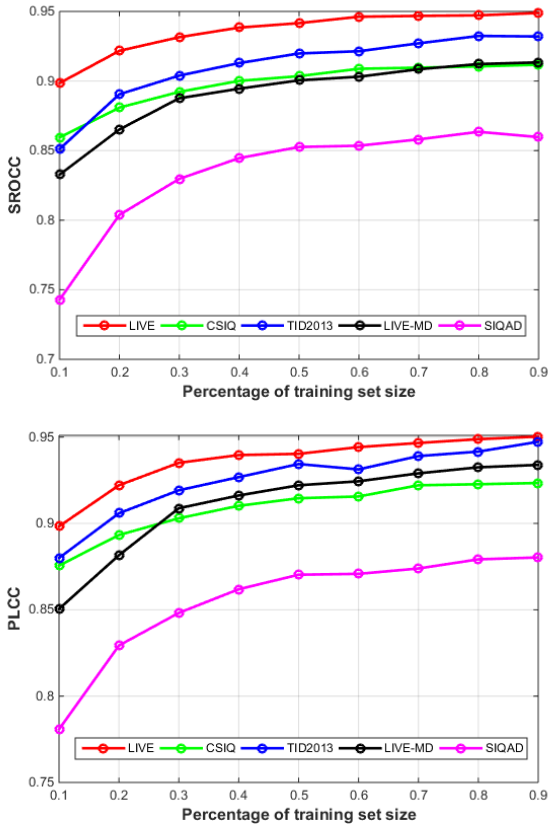


Fig. 6. Performance results for different percentages of training set size.

G. Computational Complexity

Computational complexity is another important indicator to evaluate the performance of a BIQA method because a good model is expected to provide both satisfactory efficacy and efficiency so that it can be deployed to real-world applications. The computational complexity of the competing BIQA methods is analyzed and presented in Table VI. In addition, the running time consumed by each model for estimating

TABLE VI
COMPUTATIONAL COMPLEXITY AND AVERAGE RUNNING TIME COMPARISON (IN SECONDS)

Method	Computational complexity	Running time (s)
GM-LOG	$\mathcal{O}(N(h+k))$, h: filter size, k: probability matrix size	0.0813
BRISQUE	$\mathcal{O}(Nd^2)$, d: window size	0.1024
HOSA	$\mathcal{O}(Nd^2K)$, d: window size, K: codebook size	0.4269
CORNIA-1.6K	$\mathcal{O}(Nd^2K)$, d: window size, K: codebook size	0.7848
MSDD (Pro.)	$\mathcal{O}(Nd^2K)$, d: window size, K: codebook size	1.6627
CORNIA	$\mathcal{O}(Nd^2K)$, d: window size, K: codebook size	3.2142
DIIVINE	$\mathcal{O}(Nlog(N)+m^2+N+392b)$, m: neighborhood size in DNT, b: bin number of 2D histogram	19.5607
NFREM	$\mathcal{O}(Nd^2log(N))$, d: window size of AR model	58.6212
BLIINDS-II	$\mathcal{O}((N/d^2)log(N/d^2))$, d: window size	66.3518

the quality of one 720×480 color image (taken from LIVE database) is also provided in Table VI. The experiments are performed on a personal computer with Intel(R) Core (TM) i5-6200 CPU @ 2.4 GHz and an 8GB RAM. The software platform is MATLAB R2014b. It is observed that the proposed method has a moderate time complexity. As compared to CORNIA, our method is faster and more memory-saving while without any performance loss. However, as compared to HOSA which extracts features on a 100-codeword dictionary and finally generates 14.7K-dim feature vectors, our method is slightly slower but still more memory-saving (3.2K-dim versus 14.7K-dim feature vectors). Such feature dimension reduction is rather important in the application scenarios of embedding systems or mobile devices whose memory resources are usually limited. Overall, by considering the balance between efficacy and efficiency, the performance of the method we proposed is promising and has the potential to be used to evaluate the quality of both natural images (including singly-distorted and multiply-distorted) and screen content images in practical applications.

V. CONCLUSION

This paper has presented a novel codebook-based blind image quality assessment (BIQA) method by optimizing multi-stage discriminative dictionaries (MSDDs). The unique benefits provided by the optimized MSDDs for multi-stage feature encoding (MSFE) in the context of BIQA mainly lie in the following two aspects. First, by incorporating an additional "quality-discriminative regularization" term into the traditional reconstructive error term, a unified objective function for discriminative dictionary learning is formulated. Then, this unified objective function is effectively solved by the label consistent K-SVD (LC-KSVD) algorithm, yielding a discriminative yet reconstructive dictionary. The discriminability of the learned dictionary can effectively reduce the semantic gap between the learned dictionary and the BIQA task we considered. Second, by encoding the reconstruction residuals in a stage-by-stage manner, the complementary benefits provided by the residual information for image quality prediction are also exploited and utilized. Finally, with the process of MSFE over the learned MSDDs, more discriminative feature codes can be obtained for robust quality evaluation with respect to a much smaller-size dictionary. Experimental results on several databases have demonstrated the effectiveness of the proposed method in evaluating both natural images (degraded with single distortion and multiple distortions) and screen content images. Although our method is most suitable to evaluate those common distortion types, it is still unable to offer satisfactory results on some challenge cases such as luminance change and color/contrast distortions. In the future work, we hope to seek for more complicated and effective supervised dictionary learning algorithms to simultaneously take luminance, contrast, and color components into account for blind image quality analysis.

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor and the anonymous reviewers for their constructive comments, which greatly helped in improving this paper.

REFERENCES

[1] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, "Cross-dimensional perceptual quality assessment for low bit-rate videos," *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1316-1324, Nov. 2008.

[2] L. Anegekuh, L. Sun, E. Jammeh, I.-H. Mkwawa, and E. Ifeachor, "Content-based video quality prediction for HEVC encoded videos streamed over packet networks," *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1323-1334, Aug. 2015.

[3] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "SSIM-motivated rate-distortion optimization for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 4, pp. 516-529, Apr. 2012.

[4] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "Perceptual video coding based on SSIM-inspired divisive normalization," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1418-1429, Apr. 2013.

[5] W. Gao, S. Kwong, Y. Zhou, and H. Yuan, "SSIM-based game theory approach for rate-distortion optimized intra frame CTU-level bit allocation," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 988-999, Jun. 2016.

[6] H. Liang and D. S. Weller, "Comparison-based image quality assessment for selecting image restoration parameters," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5118-5130, Nov. 2016.

[7] K. Ma, H. Yeganeh, K. Zeng, and Z. Wang, "High dynamic range image compression by optimizing tone mapped image quality index," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 3086-3097, Oct. 2015.

[8] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98-117, Jan. 2009.

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

[10] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, Article no. 011006, 2010.

[11] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935-949, Oct. 2011.

[12] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1500-1512, Apr. 2012.

[13] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment" *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378-2386, Aug. 2011.

[14] H. -W. Chang, H. Yang, Y. Gan, and M. -H. Wang, "Sparse feature fidelity for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 4007-4018, Oct. 2013.

[15] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270-4281, Oct. 2014.

[16] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684-695, Feb. 2014.

[17] U. Engelke, M. Kusuma, H.-J. Zepernick, and M. Caldera, "Reduced-reference metric design for objective perceptual quality assessment in wireless imaging," *Signal Processing: Image Communication*, vol. 24, no. 7, pp. 525-547, 2009.

[18] J. Redi, P. Gastaldo, I. Heynderickx, and R. Zunino, "Color distribution information for the reduced-reference assessment of perceived image quality," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20 no. 12, pp. 1757-1769, Dec. 2010.

[19] L. Ma, S. Li, F. Zhang, and K. N. Ngan, "Reduced-reference image quality assessment using reorganized DCT-based image representation," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 824-829, Aug. 2011.

[20] D. Liu, Y. Xu, Y. Quan, and P. Le Callet, "Reduced reference image quality assessment using regularity of phase congruency," *Signal Processing: Image Communication*, vol. 29, no. 8, pp. 844-855, 2014.

[21] F. Shao, W. Lin, S. Gu, G. Jiang, and T. Srikanthan, "Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1940-1953, May 2013.

[22] F. Shao, K. Li, W. Lin, G. Jiang, M. Yu, and Q. Dai, "Full-reference quality assessment of stereoscopic images by learning binocular receptive field properties," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 2971-2983, Oct. 2015.

[23] Q. Jiang, F. Shao, W. Lin, and G. Jiang, "Learning sparse representation for objective image retargeting quality assessment," *IEEE Transactions on Cybernetics*, DOI: 10.1109/TCYB.2017.2690452, 2017.

[24] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang, "Saliency-guided quality assessment of screen content images," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1098-1110, Jun. 2017.

[25] K. Gu, S. Wang, G. Zhai, S. Ma, X. Yang, W. Lin, W. Zhang, and W. Gao, "Blind quality assessment of tone-mapped images via analysis of information, naturalness and structure," *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 432-443, Mar. 2016.

[26] G. Yue, C. Hou, K. Gu, S. Mao, and W. Zhang, "Biologically inspired blind quality assessment of tone-mapped images," *IEEE Transactions on Industrial Electronics*, DOI: 10.1109/TIE.2017.2739708, 2017.

[27] R. Ferzli and L. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 717-728, Apr. 2009.

[28] Z. Wang, A. Bovik, and B. Evans, "Blind measurement of blocking artifacts in images," in *IEEE International Conference on Image Processing*, 2000, pp. 981-984.

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

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

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

[32] I. Kiran, T. Guha, and G. Pandey, "Blind image quality assessment using subspace alignment," in *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*, ACM, 2016.

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

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

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

[36] J. Shen, Q. Li, and G. Erlebacher, "Hybrid no-reference natural image quality assessment of noisy, blurry, JPEG2000, and JPEG images," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2089-2098, Aug. 2011.

[37] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4850-4862, Nov. 2014.

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

[39] Q. Jiang, F. Shao, G. Jiang, M. Yu, and Z. Peng, "Supervised dictionary learning for blind image quality assessment using quality-constraint sparse coding," *Journal of Visual Communication and Image Representation*, vol. 33, pp. 123-133, Nov. 2015.

[40] Q. Wu, H. Li, F. Meng, K. N. Ngan, B. Luo, C. Huang, and B. Zeng, "Blind image quality assessment based on multichannel feature fusion and label transfer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 425-440, Mar. 2016.

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

[42] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[43] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444-4457, Sep. 2016.

[44] L. Zhang, Z. Gu, X. Liu, H. Li, and J. Lu, "Training quality-aware filters for no-reference image quality assessment," *IEEE Multimedia*, vol. 21, no. 4, pp. 67-75, Oct.-Dec. 2014.

[45] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision*, 2003.

[46] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[47] Z. Jiang, Z. Lin, L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2651-2664, Nov. 2013.

[48] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, Article No. 27, Apr. 2011.

[49] N. Ponomarenko, L. Jin, O. Ieremeiev, et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57-77, 2015.

[50] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129-137, Mar. 1982.

[51] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 1, pp. 4311-4322, Nov. 2006.

[52] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Real-time no-reference image quality assessment based on filter learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 987-994.

[53] A. Liaw, and M. Wiener, "Classification and regression by randomForest," *R news*, vol. 2, no. 3, pp. 18-22, 2002.

[54] A. Hyv?rinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, nos. 4-5, pp. 411-430, Jun. 2000.

[55] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 206-220, Feb. 2017.

[56] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791-804, Apr. 2012.

[57] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 7, pp. 1294-1309, Jul. 2009.

[58] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *IEEE International Conference on Computer Vision*, 2011.
[59] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit," *CS Technion*, vol. 40, no. 8, pp. 1-15, 2008.
[60] S. Avila, N. Thome, M. Cord, E. Valle, and A. Arajo, "Pooling in image representation: The visual codeword point of view," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 453-465, May 2013.
[61] 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*, vol. 15, no. 11, pp. 3440C3451, Nov. 2006.
[62] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *ASILOMAR*, 2012, pp. 1693-1697.
[63] H. Yang, Y. Fang, and W. Lin, "Perceptual quality assessment of screen content images," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4408-4421, Nov. 2015.
[64] VQEG. (Jun. 2000). *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment*. [Online]. Available: http://www.vqeg.org/
[65] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*. London, U.K.: Chapman & Hall, 2003.

Qiuping Jiang (S'17) is currently pursing the Ph.D. degree in Signal and Information Processing at Ningbo University, Ningbo, China. Since January 2017, he has been a joint Ph.D. student with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He has authored and co-authored over ten academic papers in refereed journals and conferences in the areas of 3-D image/video processing, visual quality assessment, and saliency detection.

Feng Shao (M'16) received the B.S. and Ph.D. degrees in Electronic Science and Technology from Zhejiang University, Hangzhou, China, in 2002 and 2007, respectively. He is currently a Professor in the Faculty of Information Science and Engineering, Ningbo University, China. He was a Visiting Scholar with the School of Computer Engineering, Nanyang Technological University, Singapore. Dr. Shao received the Excellent Young Scholar Award by National Natural Science Foundation of China (NSFC) in 2016. He has authored over 100 technical articles in refereed journals and conferences in the areas of 3-D video coding, 3-D quality assessment, and image perception.

Weisi Lin (F'16) is currently a Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include image processing, perceptual modeling, video compression, multimedia communication, and computer vision. Prof. Lin serves as an Associate Editor for the IEEE Transactions on Image Processing, the IEEE Transactions on Circuits and Systems for video Technology, and the Journal of Visual Communication and Image Representation. He served as the Lead Guest Editor for a Special Issue on Perceptual Signal Processing of the IEEE Journal of Selected Topics in Signal Processing in 2012. He is the Chair of the IEEE MMTC Special Interest Group on Quality of Experience. He is a Fellow of the IEEE and the IET, an Honorary Fellow of the Singapore Institute of Engineering Technologists, and a Chartered Engineer in U.K.

Ke Gu received the B.S. and Ph.D. degrees from Shanghai Jiao Tong University, Shanghai, China, in 2009 and 2015, respectively. From March 2016 to March 2017, He was a Research Fellow at Nanyang Technological University, Singapore. He is the Associated Editor for the IEEE Access, and is the reviewer for IEEE T-NNLS, T-IP, T-CYB, T-IE, T-MM, T-CSVT, T-BC, J-STSP, SPL, etc. He reviews more than 50 journal papers each year. His research interests include visual quality assessment, contrast enhancement, and saliency detection. Dr. Gu received the Best Paper Award of ICME2016, and received the excellent Ph.D. thesis award from the Chinese Institute of Electronics (CIE) in 2016. He is the leading special session organizer in VCIP2016 and ICIP2017.

Gangyi Jiang received the M.S. degree from Hangzhou University, Hangzhou, China, in 1992, and the Ph.D. degree from Ajou University, Korea, in 2000. He is now a Professor in the Faculty of Information Science and Engineering, Ningbo University, Ningbo, China. He has published over 100 referred papers in international journals and conferences. His research interests mainly include digital video compression, multi-view video coding, and visual perception.

Huifang Sun (F'00) graduated from Harbin Military Engineering Institute, Harbin, China, and received the Ph.D. from University of Ottawa, Canada. He was an Associate Professor in Fairleigh Dickinson University in 1990. He joined to Sarnoff Corporation in 1990 as a member of technical staff and was promoted to a Technology Leader of Digital Video Communication. In 1995, he joined Mitsubishi Electric Research Laboratories (MERL) and was promoted as Vice President and Deputy Director in 2003 and currently is a Fellow of MERL. He has co-authored two books and published more than 150 Journal and Conference papers. He holds more than 61 US patents. He obtained the Technical Achievement Award for optimization and specification of the Grand Alliance HDTV video compression algorithm in 1994 at Sarnoff Lab. He was an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology and was the Chair of Visual Processing Technical Committee of IEEE Circuits and System Society.