Neurocomputing 154 (2015) 94–109

Contents lists available at ScienceDirect

Neurocomputing

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

# No-reference image quality assessment with shearlet transform and deep neural networks

Yuming Li [a,*], Lai-Man Po [a], Xuyuan Xu [a], Litong Feng [a], Fang Yuan [a], Chun-Ho Cheung [b], Kwok-Wai Cheung [c]

[a] Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong Special Administrative Region, China
[b] Department of Information Systems, City University of Hong Kong, Kowloon, Hong Kong Special Administrative Region, China
[c] Department of Computer Science, Chu Hai College of Higher Education, Hong Kong, Hong Kong Special Administrative Region, China

ARTICLE INFO

*Article history:*
Received 29 July 2014
Received in revised form
25 September 2014
Accepted 1 December 2014
Communicated by Yongzhen Huang
Available online 15 December 2014

*Keywords:*
No-reference image quality assessment
Shearlet transform
Stacked autoencoders
Softmax classification

ABSTRACT

Nowadays, Deep Neural Networks have been applied to many applications (such as classification, denoising and inpainting) and achieved impressive performance. However, most of these works pay much attention to describe how to construct the relative framework but ignore to provide a clear and intuitive understanding of why their framework performs so well. In this paper, we present a general-purpose no-reference (NR) image quality assessment (IQA) framework based on deep neural network and give insight into the operation of this network. In this NR-IQA framework, simple features are extracted by a new multiscale directional transform (shearlet transform) and the sum of subband coefficient amplitudes (SSCA) is utilized as primary features to describe the behavior of natural images and distorted images. Then, stacked autoencoders are applied as 'evolution process' to 'amplify' the primary features and make them more discriminative. Finally, by translating the NR-IQA problem into classification problem, the differences of evolved features are identified by softmax classifier. Moreover, we have also incorporated some visualization techniques to analysis and visualize this NR-IQA framework. The resulting algorithm, which we name SESANIA (ShEarlet and Stacked Autoencoders based No-reference Image quality Assessment) is tested on several database (LIVE, Multiply Distorted LIVE and TID2008) individually and combined together. Experimental results demonstrate the excellent performance of SESANIA, and we also give intuitive explanations of how it works and why it works well. In addition, SESANIA is extended to estimate quality in local regions. Further experiments demonstrate the local quality estimation ability of SESANIA on images with local distortions.



## 1. Introduction

Visual quality measurement is a vital yet complex work in many image and video processing applications. According to the dependency of reference images, the objective image quality assessment (IQA) methods are divided into three types: full-reference (FR), reduced-reference (RR) and no-reference (NR). In FR-IQA and RR-IQA methods, the whole reference images or partial information of the reference images are assumed to be available. Since information about original image are available, state-of-the-art FR-IQA methods, such as IFC [1], VIF [2] and FSIM [3], can achieve a very high correlation with human perception. However, in many practical applications the availability of the full or partial reference image's information may be very expensive or even impossible. Because of these drawbacks, NR-IQA (or blind IQA) method has recently received a great deal of attention.

Most of the conventional NR-IQA algorithms can be classified into three types as (1) Distortion-specific, (2) Natural scene statistics (NSS), and (3) Training-based. For the first type, the distortion-specific based NR-IQA algorithms usually calibrate some specific distortions, such as JPEG [4], JPEG2000 [5]. Since this kind of NR-IQA method usually implies some prior information about distortions, it is very hard to generalize them to other new distortion types. For the second type of NSS based approaches, these NR-IQAs depend on the fact that natural scenes belong to a small set in the space of all possible signals and most distortions that are prevalent in image/video processing systems destroy the specific features of natural scenes. Resent works about this type algorithms focused on developing advanced statistical models to describe the properties of natural images, and then the blind measurement of NR-IQA is achieved by measuring the variation in terms of NSS. For example, Lu et al. [6] improved the NSS model by contourlets. Moorthy et al.

* Corresponding author.
*E-mail address:* yumingli4-c@my.cityu.edu.hk (Y. Li).

http://dx.doi.org/10.1016/j.neucom.2014.12.015
0925-2312/

[7] proposed BIQI, which extracted features of NSS in the wavelet domain. Saad et al. [8] proposed BLIINDS-II that applied a NSS model of discrete cosine transform coefficients. Mittal et al. [9] proposed BRISQUE that promotes extracting NSS features from the spatial domain. For the third type of training-based NR-IQA algorithms, they usually rely on a number of features extracted from images. In which a regression model is learned based on these features and labels to predict image quality. Resent works about this type algorithms focused on using advanced machine learning methods to extract effective features to represent natural images and distorted images. For example, Li et al. [10] developed a NR-IQA algorithm using a general regression neural network. Ye et al. presented a NR-IQA framework based on unsupervised feature learning framework in [11] and a NR-IQA method based on Convolutional Neural Networks in [12].

In this paper, a new NR-IQA with use of both NSS and Training-based approaches is proposed, which is named as SESANIA (ShEarlet and Stacked Autoencoders based No-reference Image quality Assessment). The proposed algorithm is a general-purpose NR-IQA, which evaluates the image quality without incorporating any prior knowledge about distortion types. Different from our previous works [13], SESANIA does not directly use the property of NSS model in shearlet domain to construct a predictor, but utilizes the sum of subband coefficient amplitudes (SSCA) as primary features to describe the behavior of natural images and distorted images. Besides, training and learning methods are also adopted through the entire framework to achieve this new NR-IQA. The main idea of SESANIA is based on the finding that the statistical property of most natural images in shearlet domain is relatively constant. Nevertheless, distorted images usually contain more or less spread discontinuities in all directions. That is, real-world image distortions disturb the natural image statistical property and discriminate it from natural images to distorted images. Shearlets are apt at precisely detecting and locating these discontinuities or singularities. Therefore, these variations in statistical property can be easily described by shearlets and applied to describe image quality distortion.

Specifically, for natural images, the SSCA in different scales has relatively constant relationship in shearlet domain. However, this constant relationship will be disturbed if a natural image is distorted by some common distortions. Motivated by this idea, SSCA can act as a primary feature descriptor to describe an image. Thus, natural images and distorted images can be distinguished by these primary features. Recently, deep neural networks have received a great deal of attention and achieved great success on various applications, such as denoising [14,15], inpainting [15], classification [16] and natural language processing [17]. In this work, we explore applying stacked autoencoders as 'evolution process' to 'amplify' the primary features and make them more discriminative. Through this evolution process, the discriminative parts of the primary features are exaggerated. Finally, by translating the NR-IQA problem into classification problem, the differences of evolved features can be easily identified by Softmax classifier. In the implementation process, SESANIA does not incorporate any prior knowledge about distortions, which makes it suitable to many distortions and easy to extend.

The remainder of the paper is organized as follows. Section 2 introduces the detailed implementation and related techniques about SESANIA. In Section 3, experimental results and a thorough analysis of this NR-IQA framework are presented. Finally, conclusion and future works are given in Section 4.

## 2. Methodology

The proposed framework of using deep neural network for NR-IQA is illustrated in Fig. 1. The major components in this framework include: (1) SSCA extraction in shearlet domain, (2) feature evolution using stacked autoencoders, (3) evolved feature identification using softmax classifier, and (4) quality score calculation. More details will be described in the following sub-sections.

### 2.1. Shearlet transform

The proposed NR-IQA is based on the shearlet transform [18–24]. This multiscale transform is a multidimensional edition of the traditional wavelet transform [25–27], and is capable for addressing anisotropic and directional information at different scales. When the dimension $n = 2$, the affine systems with composite dilations are the collections of the form:

$$SH_\phi f(a,s,t) = <f, \quad \phi_{a,s,t}>, \quad a>0, \quad s\in R, \quad t\in R^2 \tag{1}$$

where the analyzing factor $\phi_{a,s,t}$ is called shearlet coefficient, which is defined as

$$\phi_{a,s,t}(x) = \left|\det M_{a,s}\right|^{-\frac{1}{2}}\phi(M_{a,s}^{-1}x-t) \tag{2}$$

where $M_{a,s} = B_s A_a = \begin{pmatrix} a & \sqrt{a}s \\ 0 & \sqrt{a} \end{pmatrix}$, and $A_a = \begin{pmatrix} a & 0 \\ 0 & \sqrt{a} \end{pmatrix}$, $B_s = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$. $A_a$ is the anisotropic dilation matrix and $B_s$ is the shear matrix. The analyzing functions associated to the shearlet transform are anisotropic and are defined at different scales, locations and orientations. Thus, shearlets have the ability to detect directional information and account for the geometry of multidimensional functions, which overcome the limitation of the wavelet transform.

Shearlets have a lot of very good mathematical properties [19]. For examples, shearlet is well localized (which means they are compactly supported in the frequency domain and have fast decay in the spatial domain), highly directional sensitivity and optimally sparse.

In summary, shearlets form a tight frame of well-localized waveforms, at various scales and directions, and are optimally sparse in representing images with edges. With these good properties, shearlets can provide more additional information about distorted images than the traditional wavelets and is suitable to NR-IQA work.

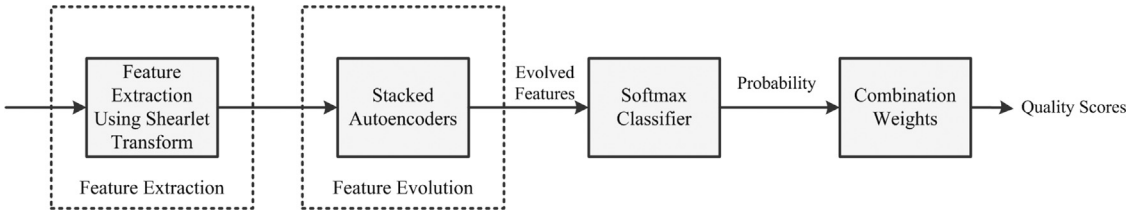



**Fig. 1.** Overview of the SESANIA framework.

### 2.2. Feature extraction

Usually, the performance of NR-IQA model is highly related to the representativeness of the features that are used for image quality prediction, which means the prediction accuracy is as good as the choice of features extracted. In this section, we will design features based on the different statistical characteristics between natural images and distorted images. Natural images indicate those images from the natural environment and they form an extremely tiny subset of all possible scenes. Natural images are not random patterns, but show a number of consistent statistical properties. One of the properties that has received considerable attention by many authors is that natural image spectra follow a power law, which is defined as

$$S(f) \approx \frac{A_s(\theta)}{f^{2-\eta(\theta)}} \tag{3}$$

where $S(f)$ denotes the image spectra, $A_s(\theta)$ is called the amplitude scaling factor, $2-\eta(\theta)$ is the frequency exponent and $\eta$ clusters around zero for natural images [28–32]. However, when natural images are distorted by some distortions, this property will be disturbed. We illustrate one instance of how the statistics of shearlet coefficients changes as a natural image is distorted by different distortions. Fig. 2 shows the shearlet coefficients of a natural image and its five distorted versions in one subband. It can be clearly seen that the process of different distortion changes the

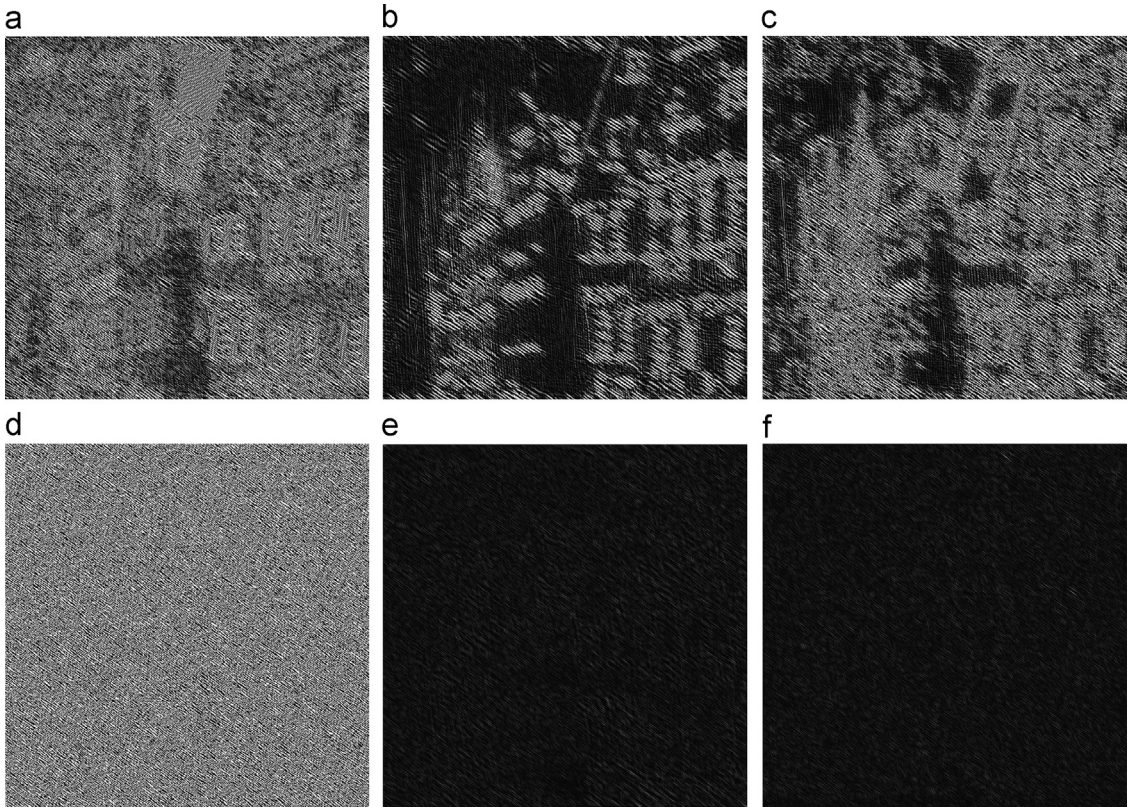



**Fig. 2.** Shearlet coefficients of a natural image and its five distorted versions in one subband. (a) Original natural image. (b) JPEG2000 compression. (c) JPEG compression. (d) Gaussian white noise. (e) Gaussian blur. (f) Rayleigh fast-fading channel simulation.

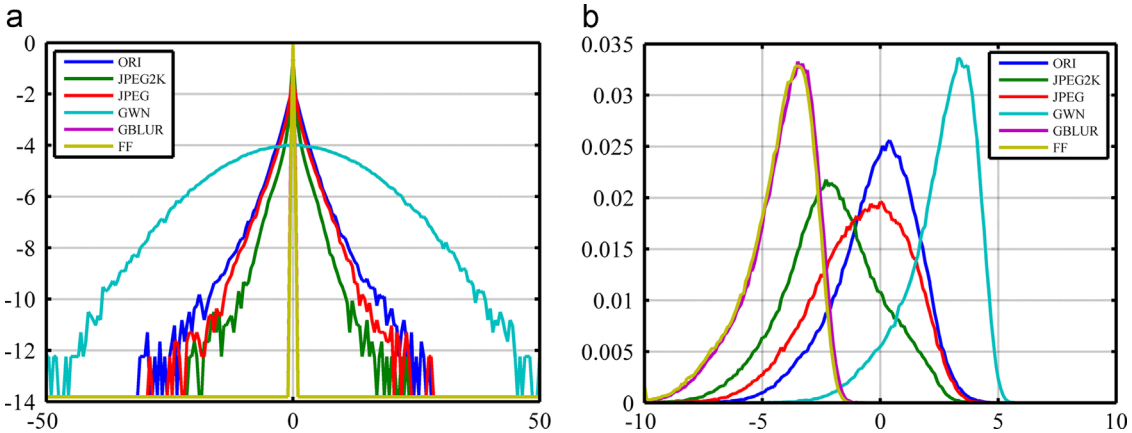



**Fig. 3.** Histograms of shearlet coefficients in Fig. 2. (a) Logarithmic histogram of original shearlet coefficients. (b) Histogram of processed shearlet coefficients ($\log_2(|coef|)$).

shearlet coefficients in different ways, which differentiates the natural images from distorted images. We further calculate the histograms of the shearlet coefficients in Fig. 2 and plot them in one coordinate. Fig. 3(a) plots the logarithmic histogram of original shearlet coefficients. We can see that blur and compression processes change the histogram into sharper peaked histogram, and noise changes it into lower peaked histogram. Fig. 3(b) shows the histogram in a different way, which means before calculating histogram, shearlet coefficients are preprocessed by $\log_2(|coef|)$. We can also see from Fig. 3(b) that blur and compression processes shift the histogram peak to the left side. On the contrary, noise shifts the histogram peak to the right side. Since the shearlet coefficient histograms of distorted and pristine images differ significantly, we can exploit these statistical differences in the design of features.

Inspired by several feature extraction methods, such as histograms of oriented gradients (HOG) [33] and its extension histograms of shearlet coefficients (HSC) [34], we can also use normalized sum of subband coefficient amplitudes (SSCA) to express this property in the shearlet domain, which is defined as

$$PF(a,s) = \frac{\sum_t |SH_\phi f(a,s,t)|}{\max\left(\sum_t |SH_\phi f(a,s,t)|\right)} \quad (4)$$

where, $SH_\phi f(a,s,t)$ is the shearlet transform of an image and $a$ is scale parameter, $s$ is direction parameter and $t$ is time shift.

Fig. 4 plots the mean SSCA for grayscale images in logarithmic coordinates (log2 of). Fig. 4 is generated by all the 29 original images and their associated distorted versions in laboratory for image and video engineering (LIVE) database [35] (detailed information about LIVE will be given in Section 3). Distortions in LIVE include JPEG2000, JPEG, Gaussian blur (GB), fast fading (FF) and Gaussian white noise (GWN). In order to provide statistical results, every original image and distorted image are randomly sampled several times by the size of 256×256. Totally 12,000 sampled blocks are obtained and 2000 for each type. Shearlet transform with 4 scales and 6 directions for each scale is applied to each of the sampled blocks, and SSCA is calculated. The horizontal axis in Fig. 4 indicates the number of subbands and each scale is divided by the gray dashed line. The vertical axis represents the mean of SSCA of the 2000 sampled blocks for each type in logarithmic coordinates.

It can be seen from Fig. 4 that common distortions disturb image statistics and make statistical property vary from that of natural images in shearlet domain. SSCA in fine scales (subbands from 7 to 24) are affected by distortions and become unnatural. SSCA in coarse scales (subbands from 1 to 6) are still less affected. Because of blurring, ringing and blocking artifact existed in JPEG2000, JPEG, GB and FF, fine scale coefficients decrease, and average energy of distorted images in fine scales become smaller than the original image, which is reflected as the decreasing of SSCA in fine scales. On the contrary, for GWN, because much high frequency components are added, the average energy of distorted image become larger than the original image, which is reflected as the increasing of SSCA in fine scales. Besides, SSCA in fine scales increase or decrease monotonously with the reduction in image quality.

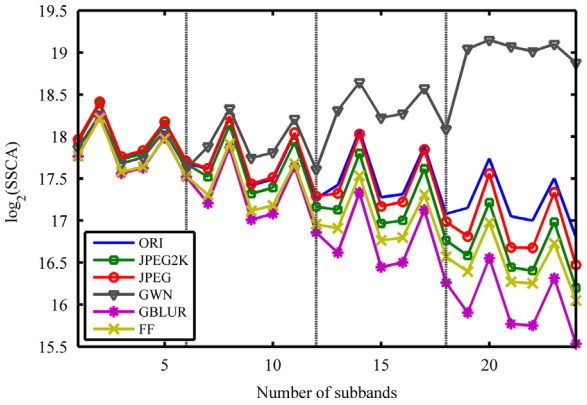



**Fig. 4.** Mean SSCA versus subband enumeration index for natural images and different distorted images in LIVE IQA database. ORI: original natural image. JPEG2K: JPEG2000 compression. JPEG: JPEG compression. GWN: Gaussian white noise. GBLUR: Gaussian blur. FF: Rayleigh fast-fading channel simulation.

### 2.3. Feature evolution

The SSCA extracted from images can serve as simple features to distinguish natural images and distorted images. However, an intuitive idea is before sending this primary feature into classifier, whether we can design a system to 'amplify' the difference between natural image features and distorted image features. Recent works about deep neural networks [36–40] provide us some ideas to solve this problem. In this paper, we propose to use stacked autoencoders to serve as an amplifier to increase the distance between natural image features and distorted image features and make them more discriminative. A stacked autoencoder is a neural network consisting of multiple layers of sparse autoencoders in which the outputs of each layer is wired to the inputs of the successive layer. Usually, two steps are implemented to obtain good parameters for a stacked autoencoder. First, each layer is treated as a sparse autoencoder and is trained individually to predetermine the encoding weights. Second, after complete the predetermine process, the decoders in each layer is discarded and fine-tuning with backpropagation are used to tune the parameters of all layers. The training process about stacked autoencoders can be found in [41].

In this work, feature extraction and feature evolution process are shown in Fig. 5. In this process, the primary features are represented by a vector which contains SSCA from RGB channel with normalization. The primary features are evolved in stacked autoencoders and the final evolved features are sent to Softmax classifier (which will be discussed in next section). To visualize the variation of features in each layer, after predetermine process, the decoders can be used as 'reconstructors' to reproduce the previous layer features using features in other layers behind this layer. Thus, both forward and backward directions of the feature evolution process are visualized.

### 2.4. Classification and quality evaluation

In the last sub-section, we apply the stacked autoencoders as an amplifier to get the evolved features and the evolved features are more discriminative. In this sub-section, we translate the NR-IQA into classification problem and use Softmax classifier to identify the image quality through final evolved features. In order to create labeled data to train the Softmax classifier and fine-tuning the stacked autoencoder, we can first classify images into several different classes based on their MOS (Mean Opinion Score). The class labels are created based on Table 1. Even though we create labels for images, our goal is not an absolute classification, but to utilize the output probability of the classification as an indication of the amount of each score in the combining weights. Therefore, there are several ways to create image labels according to how many classes we want to create. The interval boundaries in Table 1 are based on the cumulative histogram of MOS in LIVE database. We make sure that the number of images in each interval is roughly equal.

Now, we have a training set $\{(x^{(1)},y^{(1)}), \ldots, (x^{(m)},y^{(m)})\}$ of $m$ labeled data, where the input features are $x^{(i)}$, which is the final

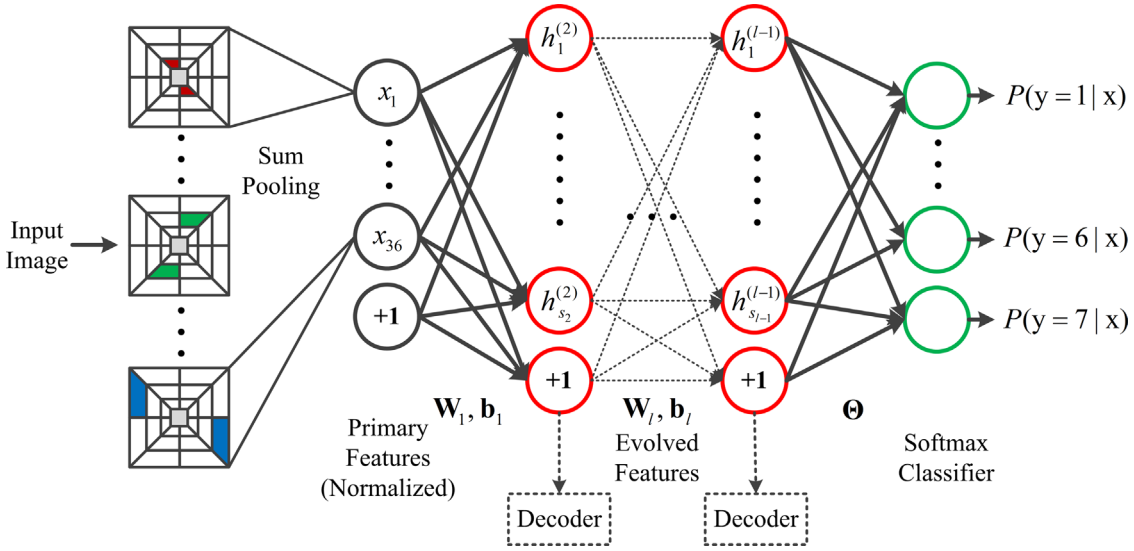



**Fig. 5.** Feature extraction and evolution process. Primary features are concatenated SSCA from RGB channels and normalization is performed after concatenation. The final evolved features are sent to Softmax classifier (Green circle). The red circle indicates sigmoid function. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 1**
The relationship between image MOS and its label.

| Class number | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **2** | **MOS** | **< 65** | **> =65** | | | | | |
| | **Class** | 1 | 2 | | | | | |
| **3** | **MOS** | **< 45** | **45–75** | **> 75** | | | | |
| | **Class** | 1 | 2 | 3 | | | | |
| **4** | **MOS** | **< 45** | **45–64** | **65–80** | **> 80** | | | |
| | **Class** | 1 | 2 | 3 | 4 | | | |
| **5** | **MOS** | **< 35** | **35–54** | **55–74** | **75–80** | **> 80** | | |
| | **Class** | 1 | 2 | 3 | 4 | 5 | | |
| **6** | **MOS** | **< 35** | **35–44** | **45–64** | **65–74** | **74–80** | **> 80** | |
| | **Class** | 1 | 2 | 3 | 4 | 5 | 6 | |
| **7** | **MOS** | **< 25** | **25–34** | **35–44** | **45–54** | **55–64** | **65–74** | **> =75** |
| | **Class** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

evolved features from the last layer of stacked autoencoder. $y^{(i)} \in \{1, 2, \ldots, K\}$ is the label of a training image.

Given an image feature $x^{(i)}$, we want to use Softmax classifier to predict the probability that $p(y^{(i)} = k|x^{(i)})$ for each value of $k = 1, \ldots, K$. Thus the output of a Softmax classifier in this problem is a $K$ dimensional probability vector, which is defined as

$$P_\theta(x^{(i)}) = \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{K} e^{\theta_l^T x^{(i)}}} \quad (5)$$

where $\theta$ are the parameters of Softmax classifier, which can be obtained by training dataset. $K$ is the number of class. As previously mentioned that our work is not a pure classification problem, a mapping is needed to convert the hypothesis $P_\theta(x^{(i)})$ into quality scores. This mapping can be done by multiplying a combining weights $\omega \in R^K$, which is a row vector. Thus, the final quality score of an image is calculated by

$$Q_i = \omega \times P_\theta^T(x^{(i)}) \quad (6)$$

where the combining weights $\omega$ are learned by calculating the least square solution of this over determined equation upon the training set.

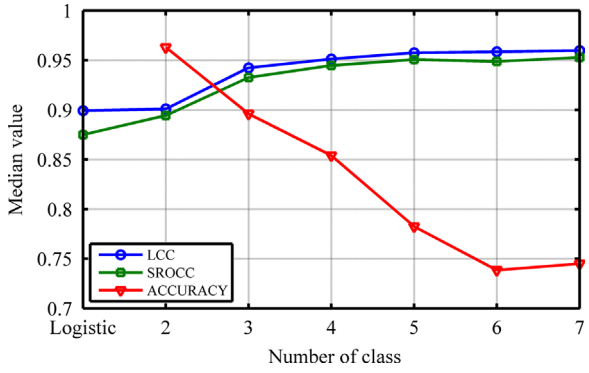



**Fig. 6.** Plot of median LCC, SROCC and classification accuracy versus number of class. Logistic means the final layer of the stacked autoencoder is a logistic regression.

Several ways for creating image labels are provided in Table 1. Now, we discuss the relationship between quality prediction performance and number of class. To describe the quality prediction performance, we choose linear correlation coefficient (LCC) scores and Spearman rank order correlation coefficient (SROCC) as the

evaluation criteria. Fig. 6 shows how LCC, SROCC and classification accuracy changes with the number of class. Fig. 6 is generated using all images in LIVE database and we report the median LCC, SROCC and classification accuracy over 100 trials. Logistic means the final layer of the stacked autoencoder is a logistic regression which outputs the normalized MOS directly. Therefore, there is no classification accuracy value for it. We can see that even through the classification accuracy decreases with the increasing of class number, the LCC and SROCC raise.

## 3. Experiments and related analysis

In order to effectively calibrate, train, test and compare the proposed NR-IQA algorithm. The following three IQA databases were used.

(1) *LIVE IQA database*. This IQA database contains 29 high-resolution 24-bits/pixel RGB original images distorted using five types of distortions at different distortion levels. These original images are distorted using the following distortion types: JPEG2000, JPEG, white Gaussian noise in the RGB components, Gaussian blur in the RGB components, and bit errors in JPEG2000 bit stream when transmitted over a simulated fast-fading Rayleigh channel. Besides, mean opinion score (MOS) and the standard deviation between subjective scores were computed for each image. MOS for LIVE is in the range 0 to 100. Higher MOS indicates higher image quality.

(2) *LIVE multiply distorted (MLIVE) IQA database* [42]. This IQA database extends one type distorted images to two types of multiply distorted images. A subjective study was conducted on 15 natural images and their distorted versions. This study was conducted in two parts to obtain human judgments on images corrupted under two multiple distortion scenarios: (a) Image storage where images are first blurred and then compressed by a JPEG encoder. (b) Camera image acquisition process where images are first blurred due to narrow depth of field or other defocus and then corrupted by white Gaussian noise to simulate sensor noise. Differential mean opinion score (DMOS) associated with distorted images are given, which is in the range 0 to 100. Different from MOS, lower DMOS indicates higher image quality.

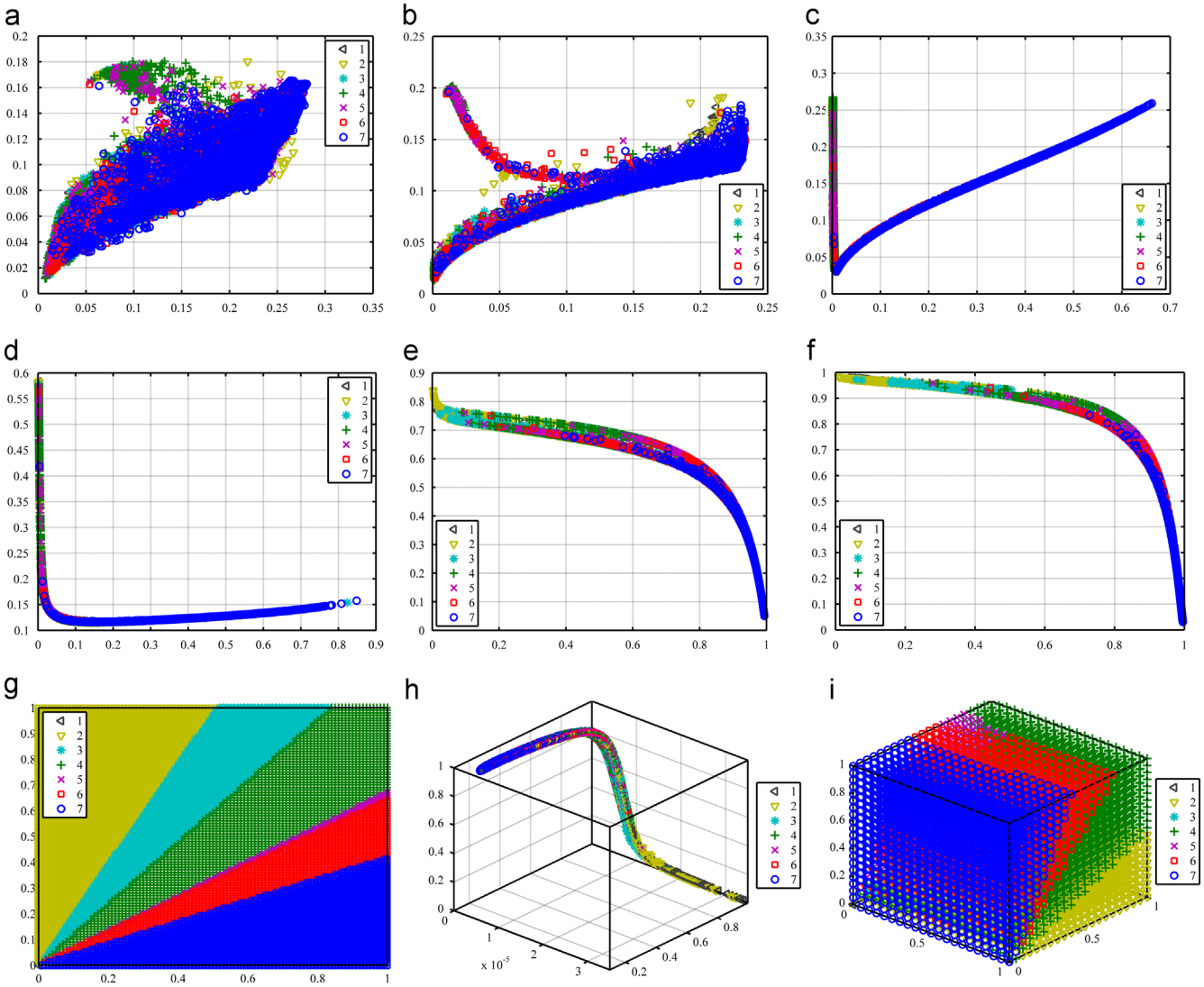



**Fig. 7.** Visualization of fine tuning process. (a) Features at first epoch. (b) Features at 15 epochs. (c) Features at 60 epochs. (d) Features at 125 epochs. (e) Features at 250 epochs. (f) Features at 400 epochs. (g) Softmax classifier boundaries after 400 iterations. (h) 3D view of features after 400 iterations. (i) 3D view of Softmax classifier boundaries. For (a)–(g), the neutron number of last layer is 2. For (h) and (i), the neutron number of last layer is 3.

(3) *TID2008 database* [43,44]. This IQA database contains 25 reference images and 1700 distorted images. These 1700 distorted images are obtained using 17 types of distortions for each reference image and every distortion has 4 levels. Mean Opinion Scores (MOS) for this database were computed for each image, which is in the range 0 to 9. Higher value indicates higher visual quality of the image. The 17 types of distortions include: Additive Gaussian noise (WN), Additive noise in color components (WNC), Spatially correlated noise (SCN), Masked noise (MN), High frequency noise (HFN), Impulse noise (IN), Quantization noise (QN), Gaussian blur (GB), Image denoising (IDN), JPEG compression (JPEG), JPEG2000 compression (JP2K), JPEG transmission errors (JPEGTE), JPEG2000 transmission errors (JP2KTE), Non eccentricity pattern noise (NEPN), Local

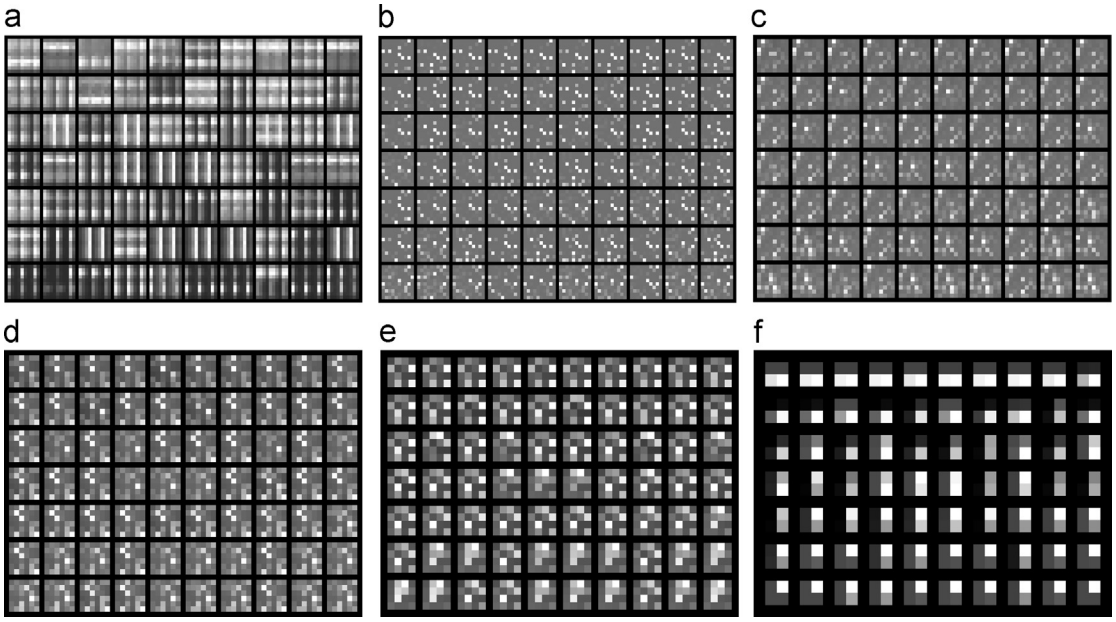



**Fig. 8.** Visualization of feature evolution process. (a) Primary features. (b) Layer 100 output features. (c) Layer 64 output features. (d) Layer 36 output features. (e) Layer 16 output features. (f) Layer 4 output features.

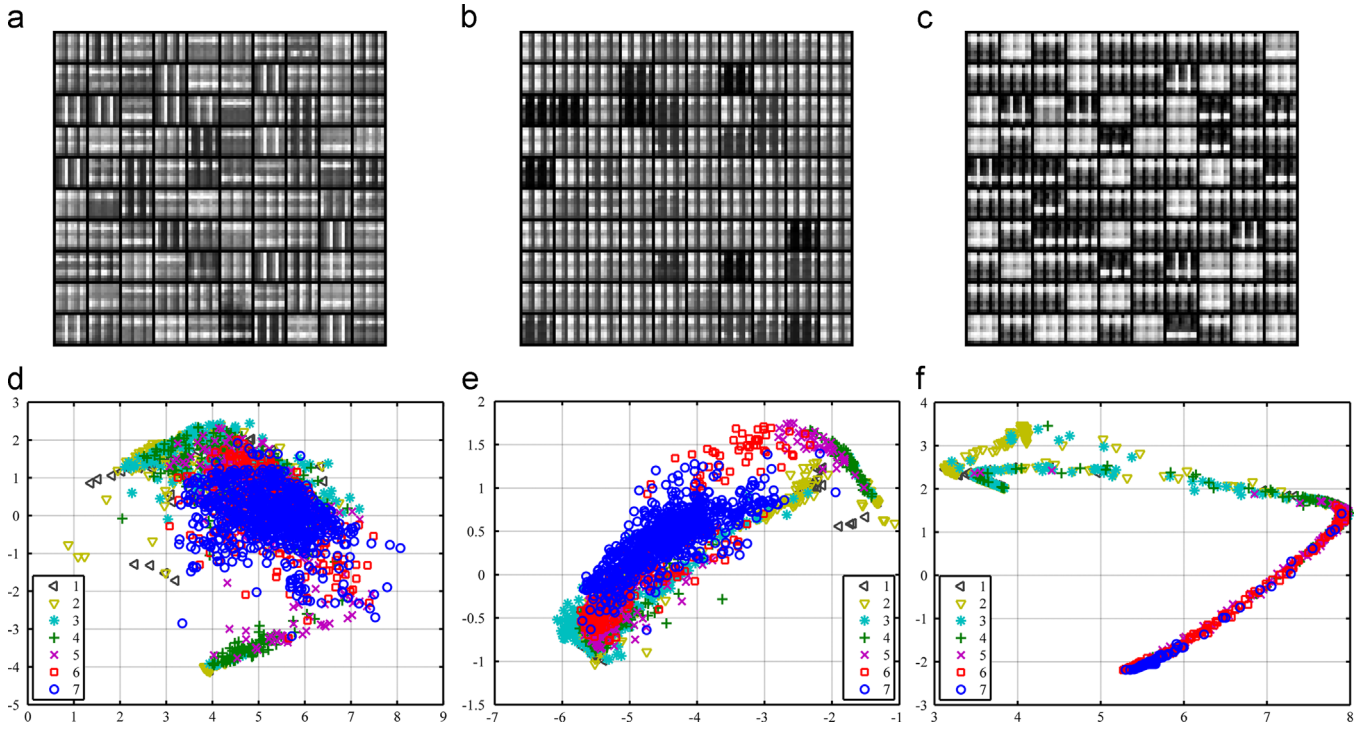



**Fig. 9.** Visualization of primary features and reconstructed primary features. (a) Primary features. (b) Reconstructed primary features using layer 49. (c) Reconstructed primary features using layer 4. (d) Two principle components of primary features. (e) Two principle components of reconstructed primary features using layer 49. (f) Two principle components of reconstructed primary features using layer 4. Parts (a)–(c) show the randomly selected 100 test features. Parts (d)–(f) show all the 2000 test features.

block-wise distortions of different intensity (LBD), Intensity shift (IS) and Contrast change (CC). In this paper, we test SESANIA on the first 13 distortion types.

### 3.1. Parameters of SESANIA

Fourier based shearlet transform is applied and the RGB channel of an image is decomposed into 4 scales (exclude approximation component) and every scale has 10 directions. The number of primary features is 120. Weight decay parameter $\lambda$ is 3e−8. Sparsity parameter $\rho$ is 0.1 and weight of sparsity penalty term $\beta$ is 5. For Softmax classifier, weight decay parameter $\lambda$ is also 3e−8, and the class number is 7. Combining weights $\omega$ is learned by calculating the least square solution of overdetermined equation upon the training subset. These parameters are fixed in the following experiments.

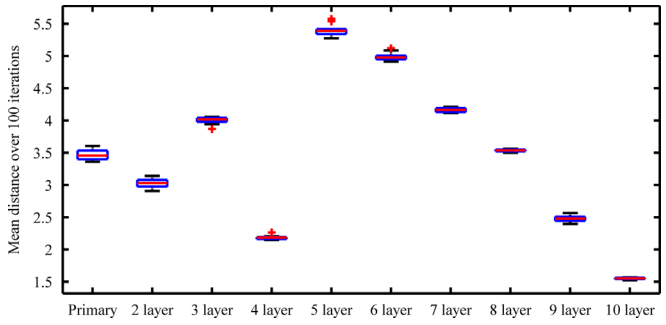



**Fig. 10.** Box plot of mean distance over 100 iterations for 9 different autoencoders and primary features.

#### 3.1.1. Forward direction visualization

In this section, we will present some intuitive visualization about how primary features changes in the deep neural network. A total of 23 natural images and their distorted versions in LIVE database are randomly selected as the training set. The remaining as the test set. Our network is trained on randomly sampled 256×256 patches taken from large images. Since the training images in LIVE have homogeneous distortions, we can assign each patch the same quality score as its source image's MOS [12]. We have sampled 2000 patches for each distortion on the training set and totally 10,000 patches are obtained as training data. 400 patches are sampled for each distortion on the test set and totally 2000 test patches are obtained.

### 3.2. Visualization of fine tuning process

In deep neural network, fine tuning strategy is usually adopted to greatly improve the performance of a stacked autoencoder. We first constructed a 120-100-81-64-49-36-25-16-9-2 autoencoder to provide an intuitive visualization of how features change in the fine tuning process. Fig. 7(a)–(f) visualizes the progression of training features during fine tuning process. It can be clear seen that with the increasing of iteration times, features are becoming much more distinctive and easier to separate. At the beginning of the iteration, we can see in Fig. 7(a) that features representing different quality images are mixed together, which is very hard to separate by a linear classifier. However, after several iterations, features indicating the same quality images are clustered and they are distributed along a quadrant arc (shown in Fig. 7(f)). Thus, a linear classifier can easily classify these features into different quality groups. Fig. 7(g) shows the Softmax classifier boundaries after 400 iterations. We can see that Softmax classifier divides the

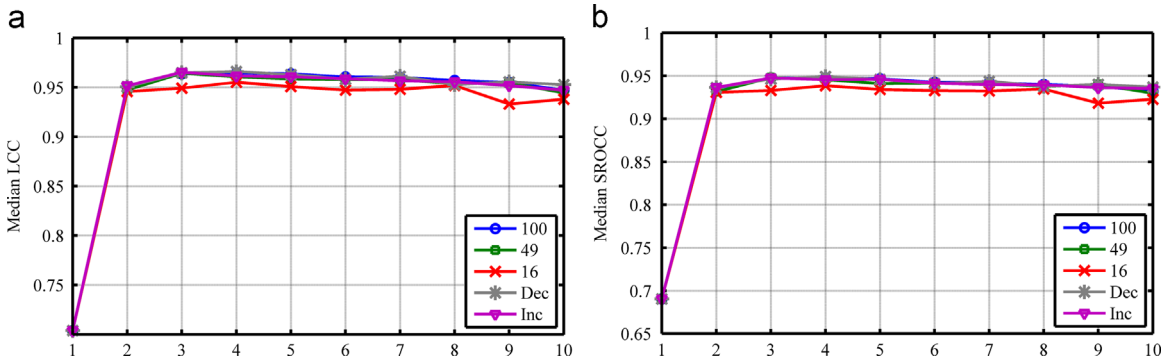



**Fig. 11.** Median LCC and SROCC over 100 iterations for 9 different autoencoders and primary features from 5 groups. (a) Median LCC. (b) Median SROCC. Horizontal axis means the number of layers.

**Table 2**
Median LCC and SROCC correlations for 1000 iterations of experiments on the three databases individually and combined together. (*Italicized* algorithms are NR-IQA algorithms.)

| | LCC | | | | | SROCC | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **LIVE** | **MLIVE** | **TID2008** | **L+T** | **L+M+T** | **LIVE** | **MLIVE** | **TID2008** | **L+T** | **L+M+T** |
| PSNR | 0.8069 | 0.7967 | 0.7633 | 0.7860 | 0.7864 | 0.8069 | 0.7257 | 0.7831 | 0.8321 | 0.8237 |
| SSIM | 0.8002 | 0.7138 | 0.6425 | 0.7574 | 0.7469 | 0.9278 | 0.6983 | 0.7308 | 0.8591 | 0.8332 |
| VIF | ***0.9574*** | 0.8459 | 0.8824 | ***0.9336*** | *0.9034* | 0.9541 | ***0.8956*** | 0.8821 | 0.9415 | 0.9163 |
| FSIM | 0.7810 | 0.8383 | 0.8715 | 0.7938 | 0.7925 | ***0.9548*** | 0.8818 | ***0.9187*** | ***0.9428*** | ***0.9311*** |
| *BLIINDS-II* | 0.9164 | 0.8203 | 0.8876 | 0.8648 | 0.7673 | 0.9124 | 0.8353 | 0.8528 | 0.8661 | 0.7623 |
| *BRISQUE* | 0.9424 | ***0.8882*** | 0.8715 | 0.8770 | 0.8234 | 0.9395 | 0.8948 | 0.8611 | 0.8780 | 0.8183 |
| *SESANIA* | 0.9476 | 0.8384 | ***0.9069*** | 0.8914 | 0.8475 | 0.9340 | 0.8362 | 0.8936 | 0.8844 | 0.8449 |

space into different non-overlapping areas and distinguishes different features based on their locations in the space. We also changed the last layer neutron number into 3 and repeat the experiment again. Fig. 7(h) and (i) shows the output training features and classifier boundaries after 400 iterations. The 3D results are very similar to 2D version. Besides, we can also notice that after fine tuning, the distance between good quality features and bad quality features are maximized. In Fig. 7(f), good quality features are clustered at the bottom right corner of the space, but bad quality features are clustered at the top left corner. In this way, the distance between the best quality feature and the worst quality feature is the diagonal of the space, which is maximized distance in this space. Similar situation also occurs in 3D space.

### 3.3. Visualization of feature evolution process

Next, we use a 120-100-81-64-49-36-25-16-9-4 autoencoder to visualize the primary features and the evolved features in several layers, which are shown in Fig. 8. A total of 70 image patches are randomly selected from 2000 test patches. They are divided into 7 groups based on their MOS, and each group has 10 images. Every row in the feature image represents one group. The first row represents best quality images (Class 7) and the last row represents worst quality images (Class 1). Detailed information about these 70 images is listed in Table A1 in the Appendix. It can be seen that the primary features in Fig. 8(a) are not very distinguishable. However, after evolution, the output features in the last layer are much easier to differentiate. In Fig. 8(f), different quality features reveal different patterns and one can even differentiate the good quality features from the bad quality features by human eyes. Besides, we can also notice from Fig. 8(f) and Table A1 that images with the same quality label tend to reveal similar patterns, which are not relevant to specific distortion types. This property also demonstrates that SESANIA is a general-purpose NR-IQA method.

#### 3.3.1. Backward direction visualization

In this section, we use another line of thought to demonstrate that the evolved features are more distinguishable than the primary features. In Section 2.3, we propose to utilize the decoders as 'reconstructors' to reproduce the previous layer features using features in other layers behind this layer. In this way, output features of every hidden layer can be converted into the primary space again. We use the 120-100-81-64-49-36-25-16-9-4 autoencoder again to evolve the primary features of 2000 test patches. The

**Table 4**
Median LCC and SROCC correlations for 1000 iterations of experiments on the LIVE multiply distorted IQA database. (*Italicized* algorithm is NR-IQA algorithms.)

| | LCC | | SROCC | |
|---|---|---|---|---|
| | **Part 1 (GB+JPEG)** | **Part 2 (GB+GWN)** | **Part 1 (GB+JPEG)** | **Part 2 (GB+GWN)** |
| PSNR | 0.7813 | 0.8088 | 0.7176 | 0.7346 |
| SSIM | 0.8125 | 0.7714 | 0.7744 | 0.7338 |
| VIF | 0.8590 | 0.8472 | 0.8808 | 0.8870 |
| FSIM | 0.8340 | 0.8385 | 0.8571 | 0.8760 |
| *BLIINDS-II* | 0.8594 | 0.8593 | 0.8539 | 0.8677 |
| *BRISQUE* | ***0.9043*** | ***0.8854*** | ***0.8947*** | ***0.8896*** |
| *SESANIA (S)* | 0.8625 | 0.8639 | 0.8423 | 0.8713 |
| *SESANIA (N)* | 0.8506 | 0.8545 | 0.8567 | 0.8623 |

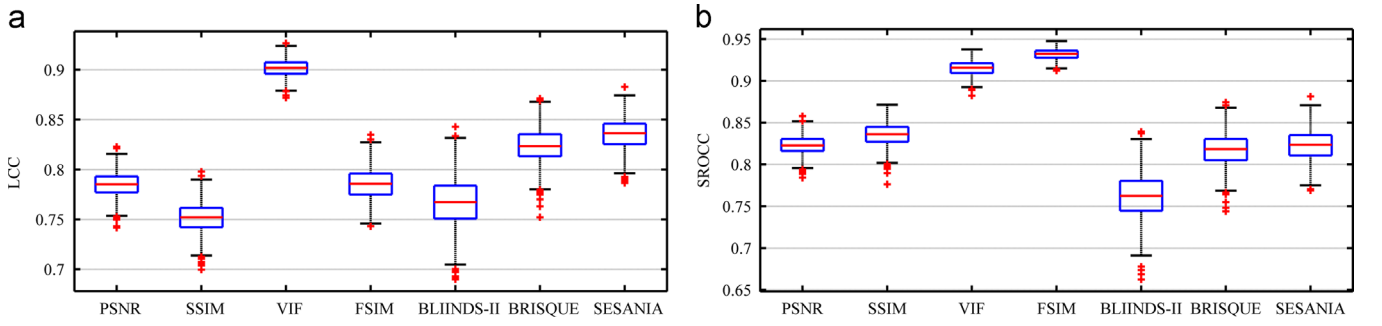



**Fig. 12.** Box plot of LCC and SROCC distributions of algorithms across 1000 trials of experiments on the combined database using LIVE, MLIVE and TID2008. (a) Box plot of LCC distribution. (b) Box plot of SROCC distribution.

**Table 3**
Median LCC and SROCC correlations for 1000 iterations of experiments on the LIVE IQA database. (*Italicized* algorithms are NR-IQA algorithms.)

| | LCC | | | | | | SROCC | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **JP2K** | **JPEG** | **GWN** | **GB** | **FF** | **ALL** | **JP2K** | **JPEG** | **GWN** | **GB** | **FF** | **ALL** |
| PSNR | 0.8669 | 0.8351 | 0.9516 | 0.8268 | 0.8665 | 0.8069 | 0.8395 | 0.8088 | 0.8838 | 0.8309 | 0.8348 | 0.8069 |
| SSIM | 0.9469 | 0.9097 | 0.9754 | 0.9077 | 0.9092 | 0.8002 | 0.9301 | 0.9712 | 0.9604 | 0.9445 | 0.9723 | 0.9278 |
| VIF | 0.9447 | 0.9692 | 0.9766 | 0.9702 | ***0.9754*** | ***0.9574*** | 0.9162 | 0.9500 | 0.9576 | 0.9709 | 0.9721 | 0.9541 |
| FSIM | 0.9034 | 0.7711 | 0.8912 | 0.8838 | 0.8316 | 0.7810 | ***0.9370*** | ***0.9721*** | 0.9574 | ***0.9804*** | ***0.9741*** | ***0.9548*** |
| *BIQI* | 0.8086 | 0.9011 | 0.9538 | 0.8293 | 0.7328 | 0.8205 | 0.7995 | 0.8914 | 0.9510 | 0.8463 | 0.7067 | 0.8195 |
| *DIIVINE* | 0.9220 | 0.9210 | ***0.9880*** | 0.9230 | 0.8680 | 0.9170 | 0.9319 | 0.9483 | ***0.9821*** | 0.9210 | 0.8714 | 0.9116 |
| *BLIINDS-II* | 0.9386 | 0.9426 | 0.9635 | 0.8994 | 0.8790 | 0.9164 | 0.9323 | 0.9331 | 0.9463 | 0.8912 | 0.8519 | 0.9124 |
| *BRISQUE* | 0.9229 | ***0.9734*** | 0.9851 | 0.9506 | 0.9030 | 0.9424 | 0.9139 | 0.9647 | 0.9786 | 0.9511 | 0.8768 | 0.9395 |
| *SESANIA(S)* | ***0.9537*** | 0.9732 | 0.9806 | ***0.9749*** | 0.9195 | 0.9476 | 0.8862 | 0.9293 | 0.9309 | 0.9410 | 0.8807 | 0.9340 |
| *SESANIA(N)* | 0.9122 | 0.9034 | 0.9587 | 0.9476 | 0.9147 | | 0.8705 | 0.8779 | 0.9362 | 0.9381 | 0.8862 | |

reconstructed primary features are obtained using every hidden layer output features. To visualize the effect of evolution process, principle components analysis (PCA) is applied to each reconstructed primary features and the first two principle components are reserved. Fig. 9(a)–(c) presents the original primary features and reconstructed primary features of randomly selected 100 test images using lay 49 and lay 4. Fig. 9(d)–(f) shows the corresponding two principle components. Compared with Fig. 9(d), in (f), principle components of features representing different image quality are separated. However, those representing same image quality are clustered together.

In addition, we also use 100-81-64-49-36-25-16-9-4 as neuron number in each hidden layer and increase one hidden layer each time to create 9 autoencoders. For example, the first autoencoder is 120-100 and has only one hidden layer 100, the second one is 120-100-81 and the last one has all the nine hidden layers. We calculate the distance of primary feature and last layer reconstructed primary feature between good quality image (Class 7) and bad quality image (Class 1 and Class 2) using Eq. (7).

$$Distance = \sqrt{(\mathbf{F}_{class7} - \mathbf{F}_{class1,2})^2} \quad (7)$$

Feature distances are obtained from 100 train-test iterations. In each iteration, we randomly select 23 natural images and their distorted versions as the training set and the remaining as the test set, and the mean distance is calculated. Training set and test set are randomly sampled as previously described. Fig. 10 shows the boxplot of the mean distance of 100 iterations for those 9 autoencoders and primary features. It can be seen that with the increasing of hidden layer, the mean distance increases first and then decreases.

### 3.3.2. *Effects of layer number and neuron number*

Several parameters are involved in the design of SESANIA. In this section, we will mainly focus on examining how layer number and neuron number in each hidden layer affect the performance of the network on the LIVE database. Other parameters are adopted as previous discussion. Similar to distance calculation in the last section, we construct five group neuron numbers and create 9 autoencoders for each group by increasing one hidden layer each time. The neuron number in each hidden layer are the same in the first three groups, they are 100, 49 and 16, respectively. We decrease the neuron number for the fourth group and increase the neuron number for the fifth group. The fourth group is 100-81-64-49-36-25-16-9-4 and the fifth group is 169-225-324-361-400-441-484-529-579. Totally 10 layers are obtained for each group including the primary features. We test each autoencoder 100

**Table 5**
Median LCC and SROCC correlations for 1000 iterations of experiments on the TID2008 database. (*Italicized* algorithm is NR-IQA algorithms.)

| | **LCC** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **WN** | **WNC** | **SCN** | **MN** | **HFN** | **IN** | **QN** | **GB** | **IDN** | **JPEG** | **JP2K** | **JPEGTE** | **JP2KTE** |
| PSNR | ***0.9454*** | ***0.9267*** | ***0.9624*** | 0.8900 | ***0.9727*** | 0.8355 | 0.9093 | 0.9376 | 0.9418 | 0.8580 | 0.8928 | 0.6796 | 0.8415 |
| SSIM | 0.7878 | 0.8073 | 0.7925 | 0.8518 | 0.8753 | 0.6991 | 0.8071 | 0.9029 | 0.9382 | 0.9362 | 0.8296 | 0.8403 | 0.7979 |
| VIF | 0.9130 | 0.9261 | 0.9064 | ***0.9393*** | 0.9583 | 0.8295 | 0.9006 | 0.9424 | 0.9475 | 0.9512 | 0.9139 | 0.8861 | 0.8450 |
| FSIM | 0.7975 | 0.8331 | 0.7929 | 0.8614 | 0.8491 | 0.6288 | 0.7942 | 0.9013 | 0.9441 | 0.9451 | ***0.9564*** | 0.8616 | 0.7730 |
| *BLIINDS-II* | 0.8234 | 0.8060 | 0.9178 | 0.7523 | 0.9504 | 0.9181 | 0.8729 | 0.9653 | 0.9597 | 0.9794 | 0.9278 | 0.8914 | 0.8920 |
| *BRISQUE* | 0.8915 | 0.8608 | 0.9228 | 0.8202 | 0.9684 | ***0.9437*** | ***0.9101*** | 0.9500 | 0.9481 | ***0.9811*** | 0.9327 | 0.8977 | 0.8800 |
| *SESANIA(S)* | 0.9144 | 0.8568 | 0.9204 | 0.8690 | 0.9079 | 0.8831 | 0.9015 | ***0.9751*** | ***0.9807*** | 0.9531 | 0.9378 | ***0.9012*** | ***0.9074*** |
| *SESANIA(N)* | 0.7868 | 0.8244 | 0.8483 | 0.8732 | 0.9187 | 0.6697 | 0.7445 | 0.9404 | 0.9054 | 0.8924 | 0.9398 | 0.8457 | 0.7256 |
| | **SROCC** | | | | | | | | | | | | |
| PSNR | 0.9038 | 0.8977 | 0.9003 | 0.8785 | 0.9238 | 0.8402 | 0.8891 | 0.9338 | 0.9233 | 0.8617 | 0.8165 | 0.7835 | 0.8150 |
| SSIM | 0.7985 | 0.8015 | 0.8003 | 0.8105 | 0.8376 | 0.6980 | 0.8876 | 0.9098 | 0.9368 | 0.9124 | 0.8617 | 0.8316 | 0.8090 |
| VIF | 0.8958 | ***0.9008*** | 0.8917 | ***0.9109*** | 0.8887 | 0.7920 | 0.8917 | 0.9534 | 0.9383 | 0.8917 | 0.9549 | 0.8541 | 0.8191 |
| FSIM | 0.8439 | 0.8406 | 0.8296 | 0.8650 | 0.8909 | 0.6977 | 0.8692 | 0.9308 | ***0.9624*** | 0.8842 | ***0.9654*** | 0.8737 | 0.8180 |
| *BLIINDS-II* | 0.7792 | 0.8071 | 0.8872 | 0.6917 | 0.9173 | 0.9083 | 0.8499 | 0.9519 | 0.9083 | ***0.9398*** | 0.9414 | 0.8838 | ***0.8767*** |
| *BRISQUE* | 0.8537 | 0.8707 | 0.8857 | 0.7575 | ***0.9248*** | ***0.9323*** | 0.8883 | 0.9289 | 0.9278 | 0.9353 | 0.9444 | ***0.8902*** | 0.8638 |
| *SESANIA(S)* | ***0.9128*** | 0.8315 | ***0.9049*** | 0.8372 | 0.8927 | 0.9101 | ***0.8951*** | ***0.9629*** | 0.9587 | 0.9311 | 0.9440 | 0.8842 | 0.8719 |
| *SESANIA(N)* | 0.8049 | 0.8327 | 0.8504 | 0.8676 | 0.9100 | 0.6907 | 0.7598 | 0.9431 | 0.8871 | 0.8712 | 0.9381 | 0.8335 | 0.7397 |

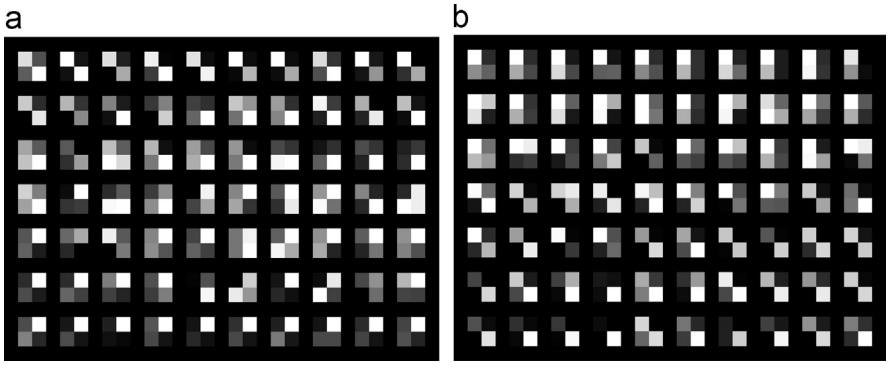



**Fig. 13.** Visualization of testing features in combined database. (a) Testing features from database mixed by LIVE and TID2008. (b) Testing features from database mixed by LIVE, MLIVE and TID2008.

times and report the median LCC scores and median SROCC scores between the predicted scores and the true MOS. When testing different autoencoder, we keep the 100 training set and test set the same. Fig. 11 shows the test results, in which first layer means primary features are sent to Softmax classifier directly without passing through evolution process. Some conclusions can be drawn from this result. (1) Feature evolution process can significantly improve the performance. (2) The proposed network is not very sensitive to neuron number in each hidden layer. (3) The proposed network is sensitive to the number of hidden layer. With the increasing of hidden layer number, the performance of the network increases first, then holds relatively constant. All these five groups achieve the best performance when the totally layer number is 3.

#### 3.3.3. Performance evaluation

In this section, we test SESANIA on LIVE, MLIVE and TID2008 database individually and combined together, and compare it with the state-of-the-art full-reference and no-reference approaches. Four FR-IQA methods include: peak-signal-to-noise ratio (PSNR), structural similarity index (SSIM) [45], visual information fidelity (VIF) and feature similarity index (FSIM). Four general purpose NR-IQA methods include: BIQI, DIIVINE [46], BLIINDS-II and BRISQUE. BIQI and DIIVINE use a representation of images in the wavelet domain and extract image features to train two stages of the algorithm. First, they train a nonlinear SVM for classification. Then, a nonlinear SVR is trained for regression within each class. In their implementation procedure, these two algorithms assume the number of distortion types is already known and do not accomplish the general-purpose IQA. Since these two algorithms are only trained and tested on LIVE database, we do not extend them to adapt to other databases. We refer to the results from Fig. 11 and construct a 120-100-81 autoencoder as the feature evolution process in the following experiments.

### 3.4. Testing on the whole database

We first test SESANIA using images of all the distortions without providing a distortion type. Five groups of experiment are conducted. The first three group experiments use these three databases individually. The fourth group experiment combined LIVE and TID2008 together and the fifth group experiment combined three databases together. For each group experiment, we randomly select 80% of reference images and their distorted versions as the training set, 20% as the test set. SESANIA is trained on randomly sampled 256×256 patches taken from large images in training set and each patch has the same quality score as its source image's MOS. One thousand randomly chosen training and test

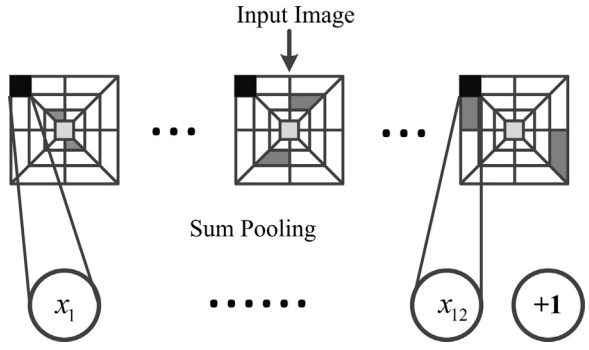



**Fig. 15.** Primary feature extraction process for local quality estimation.

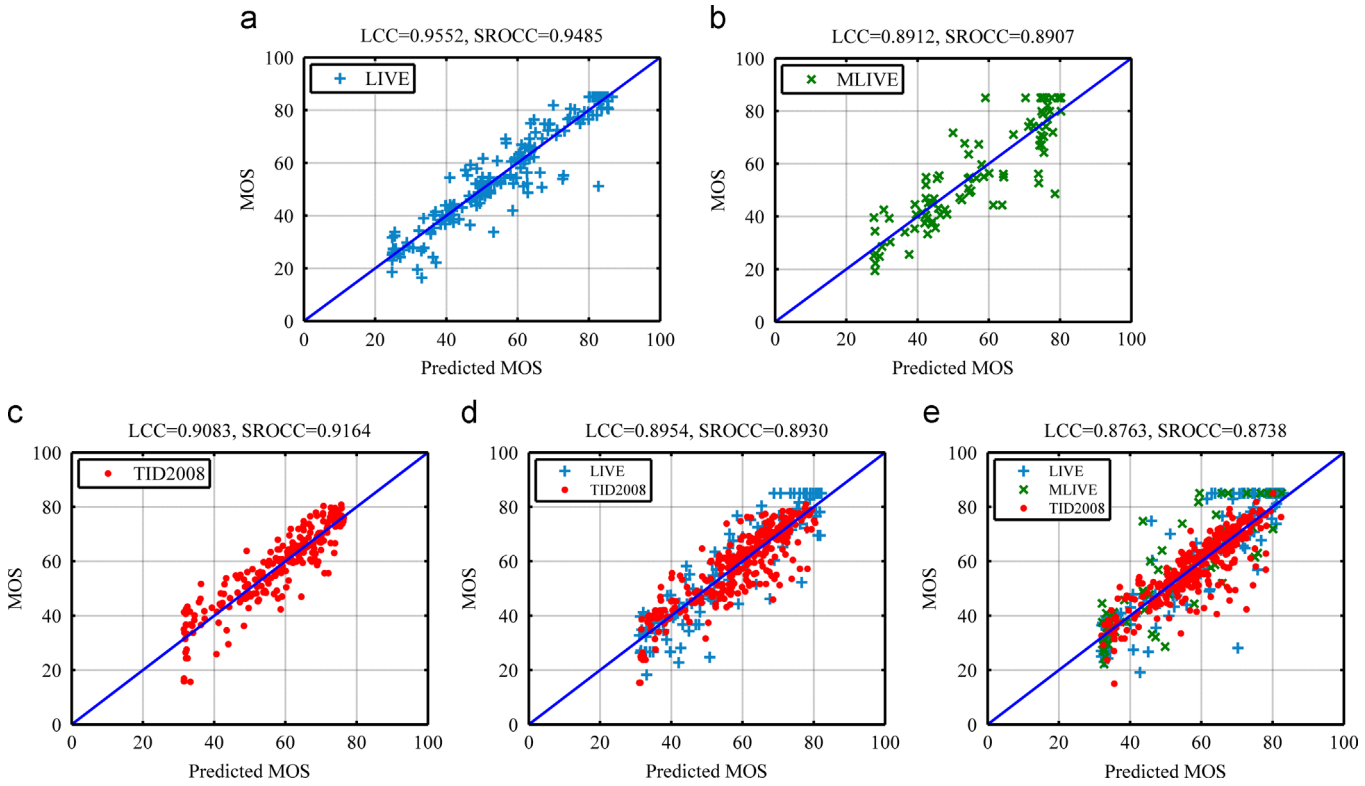



**Fig. 14.** Scatter plots of the predicted MOS versus subjective MOS on the test sets. (a) LIVE database. (b) MLIVE database. (c) TID2008 database. (d) Combined database (LIVE+TID2008). (e) Combined database (LIVE+MLIVE+TID2008).

sets were obtained, and the prediction of the quality scores was run over the 1000 iterations. To unify the MOS scores in three different databases. Subjective scores in Multiply Distorted LIVE are converted by MOS = 100 − DMOS. Subjective scores in TID2008 are converted by MOS = MOS × 10 + 15. Image labels are obtained based on Table 1 (7 classes). LCC scores and SROCC scores between

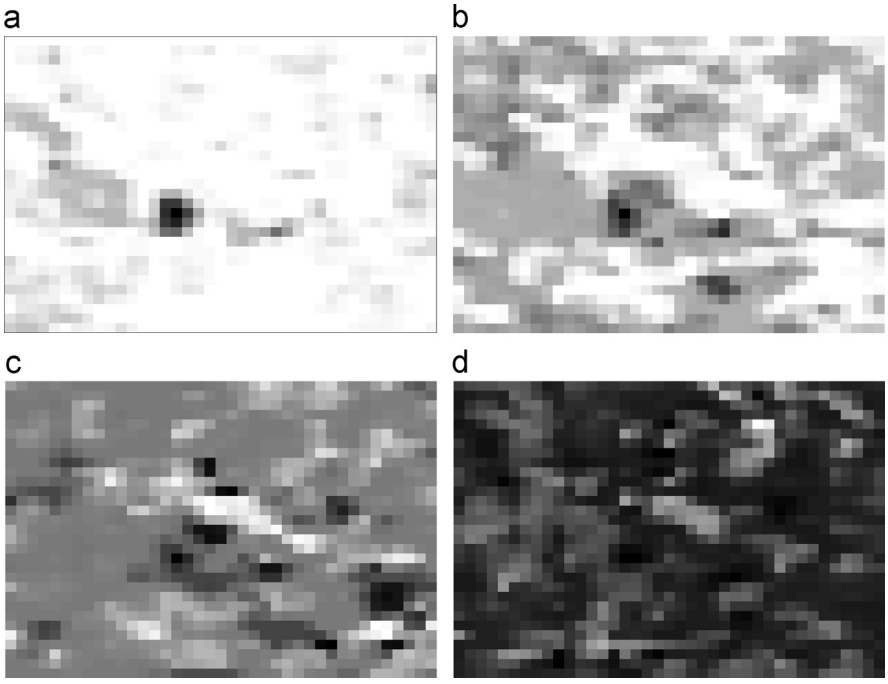



**Fig. 16.** Local quality estimation results on Building image and its GBlur distorted versions in LIVE database. (a) Original image. (b) MOS is 71.13. (c) MOS is 59.03. (d) MOS is 43.29. Brighter pixels indicate higher quality.

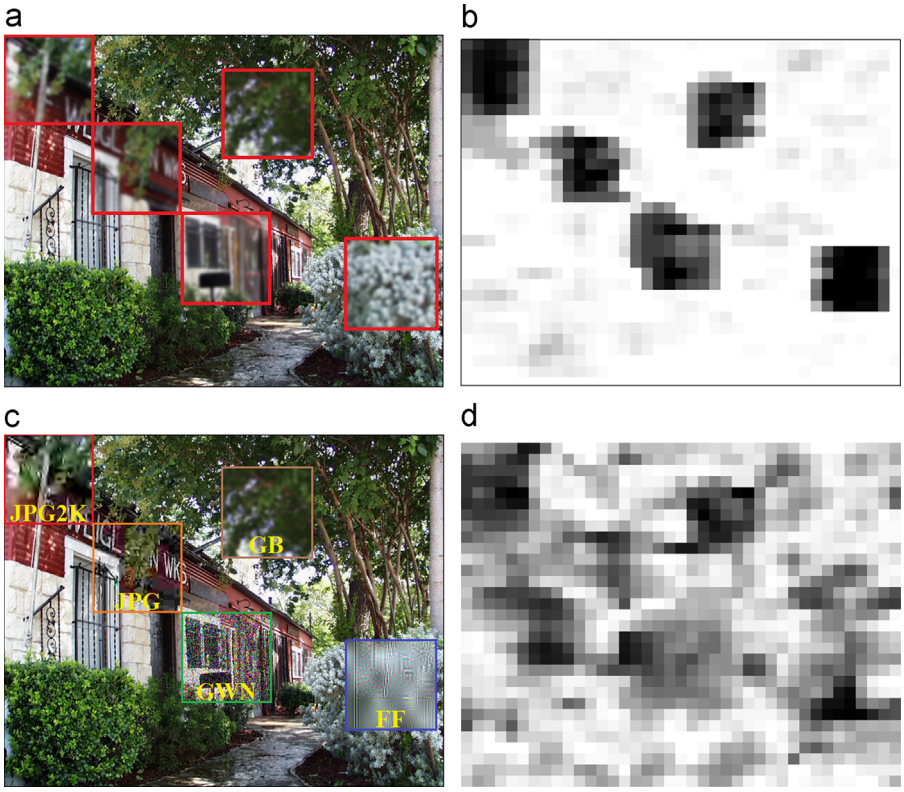



**Fig. 17.** Synthetic building images and local quality estimation results. (a) Original image contents in red boxes are replaced by its GBlur distorted versions. (b) Local quality estimation result of (a). (c) Original image contents in different color boxes are replaced by its five different distorted versions. (d) Local quality estimation result of (c). Brighter pixels indicate higher quality. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the predicted scores and the true MOS are computed for each of the 1000 iterations. For FR-IQA methods, training set are used to fit a non-linear logistic function, then testing on the test set. Table 2 reports the median LCC and SROCC over 1000 iterations. To visualize the statistical significance of the comparison, we also show box plots of the distribution of the LCC and SROCC values for each of the 1000 experimental trials, and the experiments are conducted on the three databases combined together. The plots are shown in Fig. 12(a) and (b), respectively. It can be seen that the performance of SESANIA outperforms the traditional FR-IQA approach PSNR and SSIM, and is comparable to the state-of-the-art NR-IQA and FR-IQA approaches.

### 3.5. Testing on each distortion type

Next, we conduct distortion-specific experiments and non-distortion-specific experiments on LIVE, MLIVE and TID2008. In distortion-specific experiments, we train and test on each of the distortions in the three databases. In non-distortion-specific experiments, we separate the relative test distortion set, and mix together all the other images of the three databases as training set. Testing results of three databases are listed in Tables 3–5, respectively. SESANIA(S) indicates the distortion-specific experiments and SESANIA(N) denotes the non-distortion-specific experiments. It can be seen that SESANIA achieves comparable testing results and approaches the performance of the reliable FR-IQA methods and state-of-art general purpose NR-IQA methods. For some distortion types, it also slightly outperforms these NR-IQA methods.

We also use a 120-100-81-4 autoencoder to visualize the output features of combined database. Fig. 13(a) shows the testing features from database mixed by LIVE and TID2008. Fig. 13(b) shows the testing features from database mixed by all the three databases. They are generated using the same way as Fig. 8, and their true information is also listed in Tables A2 and A3. Similar to Fig. 8(f), testing features in combined database also reveal the same pattern when they have the same quality label in spite of their distortion types. However, due to the variety of distortion types, the differences among different quality patterns in combined database are not that clear compared to Fig. 8(f). Scatter plots of the predicted MOS using SESANIA versus subjective MOS on the test sets are shown in Fig. 14. Fig. 14(a)–(c) shows the results for each individual database. Fig. 14(d)–(e) shows the results for the combined databases. In addition, the LCC and SROCC for this run are also listed on the top of each figure.

#### 3.5.1. Local quality estimation

As described in Section 2.1, shearlet is well localized in the spatial domain. Thus, SESANIA can be easily extended to detect quality of local regions. Instead of pooling all the subband coefficients, for local quality estimation purpose, every subband is divided by overlapped small blocks, and the sum pooling is conducted only on each small blocks. Fig. 15 shows the extended primary feature extraction process for local quality estimation. To simplify the calculation, we estimate local quality on gray-scale images. We construct a 40-36-25 autoencoder for the purpose of feature evolution, and the autoencoder is trained using $256\times256$ Gy-scale image patches. We scan $32\times32$ blocks with a stride of 16 for each subband.

Fig. 16 shows the local quality estimation results on Building image and its GBlur distorted versions in LIVE database. It can be clear seen that, most regions of the original image exhibit high quality (Fig. 16(a)). On the contrary, most regions of the distorted image show low quality (Fig. 16(d)). In addition, with the decreasing of image quality, local quality estimation also reveal monotonously decreasing results.

To better examine the local quality estimation power of our model, we also test it on the synthetic Building image. We first replace some of the original image contents by its GBlur distorted versions and test whether the modified model can identify the distorted regions. Fig. 17(a) shows the synthetic image and Fig. 17(b) shows the local quality estimation result. We can see that for this single distortion replacement, our model can identify the distortion regions easily. Then, we increase the difficulty and replace some of the original image contents by its five different distorted versions. Fig. 17(c) shows the synthetic image and Fig. 17(d) shows the local quality estimation result. Compared with single distortion replacement, the performance on multi-distortion replacement decreased. It is because when training the neural network using a variety of distorted image patches, its identification ability will increase. However, at the same time, it will also become more sensitive. Thus, the model will assume the local patterns on the window to be blackness distortion and give low scores on that area.

## 4. Conclusion

In this paper, a general purpose NR-IQA algorithm SESANIA is proposed, which is developed based on the shearlet transform and stacked autoencoders. We have extracted simple primary features using shearlet transform, and evolved the primary features using stacked autoencoders to make them more discriminative. By transforming the NR-IQA problem into classification problem, the difference between features can be easily distinguished by Softmax classifier. SESANIA is tested on LIVE database, LIVE Multiply Distorted database and TID2008 database individually and combined together. It generates image quality predictions well correlated with human perception, and is highly comparative to the state-of-the-art FR-IQA methods and general purpose NR-IQA algorithms. In addition, we conduct several experiments to give an intuitive visualization about the working process of SESANIA, and analyze how this algorithm works and why it works well. Furthermore, we also demonstrate that SESANIA can be easily extended to estimate quality in local regions.

Future work will involve extending the general idea of SESANIA for video-QA and stereo-QA.

## Acknowledgement

The work described in this paper was substantially supported by a grant from the City University of Hong Kong, Kowloon, Hong Kong with Project number of 7004058.

## Appendix

Image information about Figs. 8, 13(a) and (b) are provided in the following tables. The information is recorded as *Database_Distortion_ImageName_MOS*. L, M and T represent LIVE, MLIVE and TID2008, respectively.

**Table A1**
Image information of Fig. 8.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| L_FF_buildings2_85 | L_JP_sail1_78 | L_J2_woman_79 | L_J2_rapids_85 | L_GW_caps_77 | L_JP_buildings_85 | L_J2_lighthouse2_85 | L_FF_stream_85 | L_GB_dancers_85 | L_GB_sail1_77 |
| L_GB_stream_68 | L_J2_coinsinfountain_72 | L_GW_buildings2_68 | L_GB_house_74 | L_JP_sail2_66′ | L_GW_churchandcapitol_68 | L_GB_carnivaldolls_68 | L_JP_studentsculpture_73 | L_J2_churchandcapitol_65 | L_JP_monarch_71 |
| L_GB_woman_62 | L_GW_parrots_61 | L_GB_buildings2_59 | L_GW_womanhat_63 | L_FF_sail4_63 | L_GB_cemetry_58 | L_FF_statue_57 | L_GW_house_64 | L_GB_cemetry_58 | L_GW_sail4_55 |
| L_GB_sail2_54 | L_JP_studentsculpture_53 | L_FF_carnivaldolls_48 | L_GW_sail2_52 | L_GW_sail4_48 | L_GW_house_55 | L_J2_carnivaldolls_51 | L_GB_sail2_47 | L_JP_flowersonih35_49 | L_JP_flowersonih35_49 |
| L_JP_bikes_37 | L_JP_caps_40 | L_FF_sail4_42 | L_J2_statue_39 | L_GB_woman_40 | L_GB_sail1_37 | L_JP_bikes_44 | L_FF_paintedhouse_38 | L_JP_parrots_40 | L_FF_lighthouse_36 |
| L_JP_carnivaldolls_34 | L_GB_statue_34 | L_GW_womanhat_32 | L_JP_paintedhouse_26 | L_GW_monarch_34 | L_GW_caps_33 | L_GW_buildings2_28 | L_J2_studentsculpture_32 | L_GB_sail3_34 | L_GW_womanhat_32 |
| L_GB_churchandcapitol_18 | L_FF_churchandcapitol_23 | L_J2_manfishing_16 | L_J2_carnivaldolls_23 | L_GB_lighthouse_16 | L_FF_churchandcapitol_25 | L_FF_churchandcapitol_25 | L_J2_stream_23 | L_FF_churchandcapitol_25 | L_GB_monarch_24 |

**Table A2**
Image information of Fig. 13(a).

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| L_JP_house_79 | T_8_GB_womanhat_6.2 | L_JP_manfishing_79 | T_GWC_sails4_6.2 | T_HFN_door_7.0 | L_J2_plane_82 | L_J2_plane_81 | T_GWC_woman_6.6 | L_GW_statue_85 | L_FF_womanhat_85 |
| T_ID_plane_5.7 | T_JP_sails2_5.1 | T_J2_wall_5.2 | T_GB_beach_5.6 | T_GWL_caps_5.1 | T_GWL_bikes_5.0 | T_IN_sails3_5.3 | T_GWC_rapids_5.8 | T_GWL_sails2_5.2 | T_HFN_beach_5.4 |
| T_QN_flower_4.2 | T_JPTE_lighthouse1_4.4 | T_SCN_door_4.8 | T_QN_sails3_4.0 | T_GWL_rapids_5.0 | T_JPTE_wall_4.1 | T_SCN_caps_4.4 | T_HFN_parrots_4.8 | T_GWL_building_4.3 | T_GWC_ocean_4.8 |
| T_ID_flower_4.0 | L_JP_house_45 | T_SCN_caps_4.0 | T_GWL_woman_4.0 | T_ID_sails4_3.3 | L_FF_monarch_46 | L_GB_sail2_54 | T_SCN_lighthouse1_3.4 | T_QN_building_3.5 | T_SCN_caps_3.1 |
| T_ID_ocean_2.0 | T_QN_door_2.9 | L_J2_paintedhouse_43 | T_GB_sails4_2.2 | T_JP_parrots_2.3 | T_J2_door_2.9 | T_SCN_sails1_2.8 | T_JPTE_caps_2.3 | T_J2_beach_2.8 | T_GB_house_2.3 |
| T_J2_beach_1.3 | T_ID_wall_1.2 | L_GB_house_33 | L_JP_buildings2_34 | L_J2_ocean_34 | L_FF_sail1_28 | T_J2_womanhat_1.8 | L_GW_caps_33 | T_JP_house_1.4 | L_GB_house_33 |
| T_J2K_sails4_0.1 | L_GB_ocean_23 | T_J2_ocean_0.0 | T_J2_house2_0.2 | T_J2_door_0.0 | T_J2_parrots_0.2 | T_J2_beach_0.1 | T_JP_stream_0.9 | T_J2_house2_0.2 | T_J2_sails2_0.0 |

**Table A3**
Image information of Fig. 13(b).

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| L_GW_rapids_85 | L_GW_carnival-dolls_85 | L_FF_manfish-ing_85 | T_ID_light-house2_6.3 | T_GWC_statue_6.4 | L_GW_sail2_76 | L_J2_coinsinfoun-tain_83 | L_GW_bikes_85 | L_JP_house_77 | L_GW_rapids_82 |
| L_FF_flowerso-nih35_66 | T_HFN_rapids_5.4 | T_GWC_woman-hat_5.1 | T_JP_flower_5.0 | T_JP_wall_5.2 | T_HFN_flower_5.8 | M_Part2_sunrise_-noise2_73 | T_QN_door_5.7 | T__J2_house2_5.3 | T_J2TE_bikes_5.7 |
| T_IN_plane_4.0 | T_IN_door_4.2 | T_IN_house_4.3 | T_GWL_house_4.3 | T_IN_beach_4.6 | T_GB_door_4.0 | T_IN_sails2_4.1 | T_QN_beach_4.7 | L_J2_sail4_64 | T_MN_wall_4.8 |
| T_QN_par-rots_4.0 | T_JP_sails3_3.2 | T_HFN_flower_3.7 | T_QN_ocean_3.7 | L_FF_flowerso-nih35_52 | T_QN_rapids_3.0 | M_Part1_railways-tation_blur2_54 | T_J2_sails1_3.1 | T_HFN_wall_3.5 | T_J2TE_bikes_3.9 |
| M_Part1_pala-ce2_blur2_j-peg1_38 | L_JP_buildings_41 | L_J2_woman-hat_42 | T_J2_woman-hat_2.8 | M_Part1_ele-phant_blur3_j-peg1_38 | L_J2_dancers_44 | L_JP_statue_39 | T_J2TE_rapids_2.2 | T_ID_sails2_2.4 | L_JP_rapids_38 |
| L_FF_house_27 | T_JP_sails1_1.5 | M_Part1_wheel_-blur2_jpeg3_35 | L_FF_house_27 | L_JP_ocean_32 | M_Part1_sunrise_-blur3_jpeg3_27 | L_JP_rapids_29 | L_J2_cemetry_26 | L_GW_woman-hat_32 | L_GW_statue_32 |
| L_GB_rapids_20 | T_J2_house_0.0 | M_Part2_lakebuil-ding_blur3_-noise3_25 | T_J2_sails1_0.1 | L_J2_rapids_24 | L_FF_churchandca-pitol_25 | L_GB_rapids_20 | T_J2_sails4_0.1 | L_J2_carnival-dolls_23 | L_JP_churchand-capitol_20 |

## References


[1] Hamid R. Sheikh, Alan C. Bovik, Veciana Gustavo De, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Trans. Image Process. 14 (12) (2005) 2117–2128.

[2] Hamid R. Sheikh, Alan C. Bovik. A visual information fidelity approach to video quality assessment, in: The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2005, pp. 23–25.

[3] Lin Zhang, D. Zhang, Xuanqin Mou, FSIM: a feature similarity index for image quality assessment, IEEE Trans. Image Process. 20 (8) (2011) 2378–2386.

[4] Wang, Zhou, Hamid R. Sheikh, Alan C. Bovik, No-reference perceptual quality assessment of JPEG compressed images, in: Proceedings of IEEE International Conference on Image Processing 2002, vol. 1, 2002, pp. I–477.

[5] Hamid R. Sheikh, Alan C. Bovik, Lawrence Cormack, No-reference quality assessment using natural scene statistics: JPEG2000, IEEE Trans. Image Process. 14 (11) (2005) 1918–1927.

[6] Kai Zeng Lu Wen, Dacheng Tao, Yuan Yuan, Xinbo Gao, No-reference image quality assessment in contourlet domain, Neurocomputing 73 (4) (2010) 784–794.

[7] Anush Krishna Moorthy, Bovik Alan Conrad, A two-step framework for constructing blind image quality indices, IEEE Signal Process. Lett. 17 (5) (2010) 513–516.

[8] Michele A. Saad, Alan C. Bovik, Christophe Charrier, Blind image quality assessment: a natural scene statistics approach in the DCT domain, IEEE Trans. Image Process. 21 (8) (2012) 3339–3352.

[9] Anish Mittal, Anush Krishna Moorthy, Bovik Alan Conrad, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (12) (2012) 4695–4708.

[10] Chaofeng Li, Alan Conrad Bovik, Xiaojun Wu, Blind image quality assessment using a general regression neural network, IEEE Trans. Neural Networks 22 (5) (2011) 793–799.

[11] Ye, Peng, Jayant Kumar, Le Kang, David Doermann, Unsupervised feature learning framework for no-reference image quality assessment, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1098–1105.

[12] L. Kang, P. Ye, Y. Li, D. Doermann. Convolutional neural networks for no-reference image quality assessment, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[13] Yuming Li, Lai-Man Po, Xuyuan Xu, Litong Feng, No-reference image quality assessment using statistical characterization in the shearlet domain, Signal Process. Image Commun. 29 (7) (2014) 748–759.

[14] Jain, Viren, H. Sebastian Seung. Natural image denoising with convolutional networks, in: NIPS, vol. 8, pp. 769-776. 2008.

[15] Xie, Junyuan, Linli Xu, Enhong Chen. Image denoising and inpainting with deep neural networks, in: NIPS, pp. 350-358. 2012.

[16] Krizhevsky, Alex, Ilya Sutskever, Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks, in: NIPS, vol. 1, 2, 4. 2012.

[17] Socher, Richard, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, Christopher D. Manning. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection, in: NIPS, 24, pp. 801-809, 2011.

[18] Yi Sheng, Demetrio Labate, Glenn R. Easley, Hamid Krim, A shearlet approach to edge analysis and detection, IEEE Trans. Image Process. 18 (5) (2009) 929–941.

[19] Glenn Easley, Demetrio Labate, Wang-Q. Lim, Sparse directional image representations using the discrete shearlet transform, Appl. Comput. Harmon. Anal. 25 (1) (2008) 25–46.

[20] Kutyniok, Gitta, Wang-Q. Lim. Image Separation using Wavelets and Shearlets. arXiv preprint arXiv: vol. 1101.p. 0553, 2011.

[21] Gitta Kutyniok, Wang-Q. Lim, Xiaosheng Zhuang, Digital Shearlet Transforms, Shearlets, Birkhäuser, Boston (2012) 239–282.

[22] Kutyniok, Gitta, Morteza Shahram, Xiaosheng Zhuang. Shearlab: A Rational Design of a Digital Parabolic Scaling Algorithm. arXiv preprint arXiv:vol. 1106. p. 1319, 2011.

[23] Donoho, David L., Gitta Kutyniok, Morteza Shahram, Xiaosheng Zhuang, A rational design of a digital shearlet transform, in: RN (RN+ 1) vol. 8, 2, 2011 p. 1.

[24] Gitta Kutyniok, Morteza Shahram, David L. Donoho, Development of a digital shearlet transform based on pseudo-polar FFT, SPIE Optical Engineering+Applications, International Society for Optics and Photonics, 2009 (74460B-74460B).

[25] Ingrid Daubechies, Ten Lectures on Wavelets, vol. 61, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[26] Stéphane Mallat, A Wavelet Tour of Signal Processing, Access Online via Elsevier, 1999.

[27] C. Sidney Burrus, Ramesh A. Gopinath, Haitao Guo, Jan E. Odegard, Ivan W. Selesnick, Introduction to Wavelets and Wavelet Transforms: A Primer, vol. 23, Prentice Hall, Upper Saddle River, 1998.

[28] R. P. Millane, S. Alzaidi, W. H. Hsiao, Scaling and power spectra of natural images, in: Proc. Image and Vision Computing New Zealand, pp. 148-153. 2003.

[29] David J. Field, Relations between the statistics of natural images and the response properties of cortical cells, J. Opt. Soc. Am. A 4 (12) (1987) 2379–2394.

[30] D.J. Tolhurst, Y. Tadmor, Tang Chao, Amplitude spectra of natural images, Ophthal. Physiol. Opt. 12 (2) (1992) 229–232.

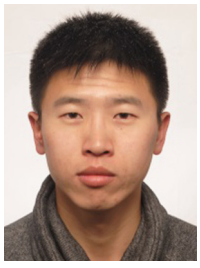
[31] Daniel L. Ruderman, Origins of scaling in natural images, Electronic Imaging: Science & Technology, International Society for Optics and Photonics (1996) 120–131.

[32] David J. Field, Nuala Brady, Visual sensitivity, blur and the sources of variability in the amplitude spectra of natural scenes, Vision Res. 37 (23) (1997) 3367–3383.
[33] Dalal, Navneet, Bill Triggs, Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR, 1, 2005, pp. 886-893.
[34] Schwartz, William Robson, Ricardo Dutra da Silva, Larry S. Davis, Helio Pedrini, A novel feature descriptor based on the shearlet transform, in: IEEE International 18th Conference on Image Processing (ICIP), 2011, pp. 1033–1036.
[35] Sheikh, Hamid R., Zhou Wang, Alan C. Bovik, L. K. Cormack, Image and Video Quality Assessment Research at LIVE. ⟨http://live.ece.utexas.edu/research/quality/⟩, 2003.
[36] Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, Yoshua Bengio, An empirical evaluation of deep architectures on problems with many factors of variation, Proceedings of the 24th International Conference on Machine Learning, ACM (2007) 473–480.
[37] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, Samy Bengio, Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. 11 (2010) 625–660.
[38] J. Ian, Goodfellow, V. Quoc Le, Andrew M. Saxe, Honglak Lee, Andrew Y. Ng, Measuring Invariances in Deep Networks, in: NIPS, 9, pp. 646-654. 2009.
[39] D. Matthew, Zeiler, Rob Fergus, Visualizing and Understanding Convolutional Neural Networks. arXiv preprint arXiv:1311.2901, 2013.
[40] Geoffrey E. Hinton, Ruslan R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[41] Andrew Ng, Jiquan Ngiam, Chuan Yu Foo, Yifan Mai, Caroline Suen, Unsupervised Feature Learning and Deep Learning. ⟨http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial⟩.
[42] Jayaraman, Dinesh, Anish Mittal, Anush K. Moorthy, Alan C. Bovik, Objective quality assessment of Mtiply distorted images, in: IEEE Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), 2012, pp. 1693–1697.
[43] Ponomarenko Nikolay, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, M. Carli, F. Battisti, TID2008-A database for evaluation of full-reference visual quality assessment metrics, Adv. Mod. Radioelectron. 10 (4) (2009) 30–45.
[44] Ponomarenko, Nikolay, Federica Battisti, Karen Egiazarian, Jaakko Astola, Vladimir Lukin, Metrics performance comparison for color image database, in: Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics, vol. 27, 2009.
[45] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, Eero P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[46] Anush Krishna Moorthy, Bovik. Alan Conrad, Blind image quality assessment: From natural scene statistics to perceptual quality, IEEE Trans. Image Process. 20 (12) (2011) 3350–3364.

**Yuming Li** (S'13) received his B.Eng. degree from Huazhong University of Science and Technology in 2011, his M.Eng. degree from Huazhong University of Science and Technology in 2013. He is currently pursuing his Ph.D. degree in City University of Hong Kong. His research interests include image and video processing, multiscale analysis and machine learning.

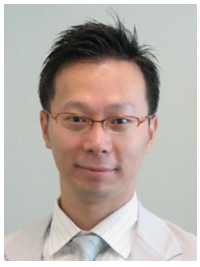
**Lai-Man Po** (M'92–SM'09) received the B.S. and Ph.D. degrees in electronic engineering from the City University of Hong Kong, Kowloon, Hong Kong, in 1988 and 1991, respectively. Since 1991, he has been with the Department of Electronic Engineering, City University of Hong Kong, and is currently an Associate Professor and lab director of TI Educational Training Centre. He has published over 140 technical journal and conference papers. His current research interests are in the areas of image and video coding with emphasis on fast encoding algorithms, new motion compensated prediction techniques and 3D video processing. He is a member of the Technical Committee on Multimedia Systems and Applications, IEEE Circuits and Systems Society. Dr. Po was the chairman of the IEEE Signal Processing Hong Kong Chapter in 2012 and 2013. He was an Associate Editor of the HKIE Transactions in 2011 to 2013. He also served on the organizing committees of the IEEE International Conference on Acoustics, Speech and Signal Processing in 2003, the IEEE International Conference on Image Processing in 2010, and other conferences.

**Xuyuan Xu** (S'11) received his B.E. in Information Engineering in 2010 in City University of Hong Kong. His final year project "Stereoscopic Video generation from Monoscopic Video" won the Best Tertiary Student Project of Asia Pacific International and Communication Awards (APICTA) in 2010. He is currently a PhD student in the Department of Electronic Engineering in City University of Hong Kong. His research interests include 3D video coding and 3D view synthesis.



**Litong Feng** (S'12) received the B.E. degree in Electronic Science and Technology from Harbin Institute of Technology, Harbin, China, in 2008 and the M.E. degree in Optical Engineering from Tianjin Jinhang Institute of Technical Physics, Tianjin, China, in 2011. He is currently working toward the Ph.D. degree in the Department of Electronic Engineering, City University of Hong Kong. His research interests include video processing for vital signs and optical system design.

**Fang Yuan** (S'14) received his B.Sc. degree in Physics from Central South University, Changsha, China, in 2009 and the M.E. degree in Communication and Information System from Sun Yat-sen University, Guangzhou, China, in 2012. He is currently working toward the Ph.D. degree in the Department of Electronic Engineering, City University of Hong Kong. His research interests include image and biomedical signal processing and machine learning.



**Chun-Ho Cheung** received his BEng (Hons) degree in computer engineering and PhD degree in electronic engineering from City University of Hong Kong, in 1996 and 2002, respectively. Currently, he is an Assistant Professor at the Department of Information Systems, City University of Hong Kong. His research interests include image coding, motion estimation and e-Learning.

**Kwok-Wai Cheung** received B.E., M.S., and Ph.D. degrees, all in electronic engineering, from City University of Hong Kong in 1990, 1994, and 2001, respectively. He was a Research Student/Research Assistant with the Department of Electronic Engineering, City University of Hong Kong, from 1996 to 2002. He joined the Chu Hai College of Higher Education, Hong Kong, in 2002. Currently, he is an Associate Professor with the Department of Computer Science, Chu Hai College of Higher Education. His current research interests include the areas of image/video coding and multimedia database.