No-Reference Video Quality Assessment With 3D Shearlet Transform and Convolutional Neural Networks

Yuming Li, *Student Member, IEEE*, Lai-Man Po, *Senior Member, IEEE*, Chun-Ho Cheung, Xuyuan Xu, *Student Member, IEEE*, Litong Feng, *Student Member, IEEE*, Fang Yuan, *Student Member, IEEE*, and Kwok-Wai Cheung

***Abstract*— In this paper, we propose an efficient general-purpose no-reference (NR) video quality assessment (VQA) framework that is based on 3D shearlet transform and convolutional neural network (CNN). Taking video blocks as input, simple and efficient primary spatiotemporal features are extracted by 3D shearlet transform, which are capable of capturing natural scene statistics properties. Then, CNN and logistic regression are concatenated to exaggerate the discriminative parts of the primary features and predict a perceptual quality score. The resulting algorithm, which we name shearlet- and CNN-based NR VQA (SACONVA), is tested on well-known VQA databases of Laboratory for Image & Video Engineering, Image & Video Processing Laboratory, and CSIQ. The testing results have demonstrated that SACONVA performs well in predicting video quality and is competitive with current state-of-the-art full-reference VQA methods and general-purpose NR-VQA algorithms. Besides, SACONVA is extended to classify different video distortion types in these three databases and achieves excellent classification accuracy. In addition, we also demonstrate that SACONVA can be directly applied in real applications such as blind video denoising.**

***Index Terms*— 3D shearlet transform, autoencoder (AE), convolutional AE (CAE), convolutional neural network (CNN), distortion identification, no-reference (NR) video quality assessment (VQA).**

I. Introduction

NOWADAYS, with the rapid development of multimedia and network technology, videos are much easier to be generated and transmitted by many different devices, and shared by many social media, such as Facebook, Twitter, YouTube, and Instagram. Since a large number of video contents are produced every day for entertainment or education of human viewers, it is of prime importance to guarantee that the perceived visual quality of these videos is still maintained at an acceptable level at the end user after the production and distribution chains. To achieve this goal, effective image and video quality assessment (I/VQA) algorithms are needed and have recently attracted considerable research attention.

I/VQA can be completed using two types of methods, which are subjective and objective I/VQA methods. Subjective I/VQA methods rely on the opinions of a large number of viewers, which makes them expensive to implement and impractical in real applications. Although subjective I/VQA methods are cumbersome in real applications, they are usually adopted to design a subjective score for each image or video in I/VQA database, such as the mean opinion score (MOS) in each I/VQA database. Objective I/VQA methods refer to designing algorithms to automatically predict the visual quality of an image or a video that is consistent with human perception. According to the dependency of reference images or videos, objective I/VQA methods are usually divided into three types: full-reference (FR), reduced-reference (RR), and no-reference (NR).

FR-I/VQA and RR-I/VQA metrics assume that the whole reference signal or partial information of the signal is available, and do a comparison between the reference signal and tested signal. Since information about the original signal can be used as reference, state-of-the-art FR-I/VQA methods can achieve a high correlation with human perception. Some state-of-the-art FR-IQA algorithms include information fidelity criterion [1], visual information fidelity (VIF) [2], and feature similarity index [3]. Prominent FR-VQA includes spatiotemporal most-apparent-distortion model (ST-MAD) [4], ViS3 [5], video quality metric [6], and motion-based video integrity evaluation [7].

NR-I/VQA metrics exploit only the tested signal and have no need of any information about reference signal. Because of this advantage, NR-I/VQA algorithms have much wider applicability and received a great deal of attention. Previous researchers have attempted to develop distortion-specific NR-I/VQA algorithms. These algorithms calibrate some specific distortions, such as Joint Photographic Experts Group (JPEG) [8], JPEG 2000 [9], and H.264/AVC [10]. Although these methods work well for the specific distortions, it is not easy

Manuscript received October 30, 2014; revised January 29, 2015, March 6, 2015, and April 17, 2015; accepted April 29, 2015. Date of publication May 6, 2015; date of current version June 2, 2016. This work was supported by the City University of Hong Kong, Hong Kong, under Project 7004058. This paper was recommended by Associate Editor A. Loui.

Y. Li, L.-M. Po, X. Xu, L. Feng, and F. Yuan are with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong (e-mail: yumingli4-c@my.cityu.edu.hk; xuyuanxu2-c@my.cityu.edu.hk; litonfeng2-c@my.cityu.edu.hk; fangyuan7-c@my.cityu.edu.hk; eelmpo@cityu.edu.hk).

C.-H. Cheung is with the Department of Information Systems, City University of Hong Kong, Hong Kong (e-mail: is.tc@cityu.edu.hk).

K.-W. Cheung is with the Department of Computer Science, Chu Hai College of Higher Education, Hong Kong (e-mail: kwcheung@chuhai.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2015.2430711

for them to be generalized to other new distortion types. Thus, these approaches are inferior to the state-of-the-art approaches. Nowadays, many researchers have paid much effort to investigate natural scene statistics (NSS)-based general-purpose NR-I/VQA algorithms. Some successful examples of such kind of NR-IQA approaches include distortion identification-based image verity and integrity evaluation index [11], BLind Image Integrity Notator (BLIINDS) using discrete cosine transform (DCT) statistics–II index [12], and blind/referenceless image spatial quality evaluator [13]. Compared with the NSS-based NR-IQA approach, nowadays, training-based NR-IQA is a new trend. With the development of feature learning methods, training-based NR-IQA approaches learn discriminative features directly from raw image patches without using hand-crafted features. These methods deal with small image patches (such as 32×32) and the whole image quality score is the average score of small patches. The representative works about this type of NR-IQA work include codebook representation for no-reference image assessment (CORNIA) [14] and convolutional neural networks for no-reference image quality assessment (CNNNRIQA) [15]. CORNIA aims at training image representation kernels directly from raw image pixels using unsupervised feature learning and CNNNRIQA integrates feature learning and regression into one optimization process using convolutional neural networks (CNNs). However, compared with general-purpose NR-IQA algorithms, there is still a lack of prominent general-purpose NR-VQA algorithms. One of the latest representative works of general-purpose NR-VQA is proposed by Saad *et al.* [16]. The authors successfully extended their previous NR-IQA idea to NR-VQA and proposed video BLIINDS. In their work, they proposed a spatiotemporal NSS model and a motion model to blindly predict video quality and achieved promising results. Besides, Xu *et al.* [17] proposed a NR-VQA method that is based on feature learning. In their work, frame-level features are first extracted by unsupervised feature learning and these features are applied to train a linear support vector regressor. Then, the final score of a single video is obtained by combining the frame-level scores using temporal pooling.

Since there exist very few general-purpose NR-VQA algorithms that have been shown to consistently correlate well with human judgments of temporal visual quality, in this paper, a new general-purpose NR-VQA algorithm with the use of 3D shearlet transform and CNN is proposed. The proposed NR-VQA algorithm, which is named shearlet- and CNN-based NR VQA (SACONVA), evaluates video quality without incorporating any prior knowledge about distortion types. Inspired by our previous general-purpose NR-IQA algorithm shearlet-based no-reference image quality assessment [18], we propose to extract simple and efficient primary spatiotemporal features from video blocks using 3D shearlet transform and these features are able to capture NSS properties. Then, these primary spatiotemporal features are further evolved by CNN. Through the evolution process, the discriminative parts of the primary features are exaggerated. Finally, logistic regression is applied to predict a perceptual quality score. The original 2D CNN is designed for capturing object features in an image, which is not suitable for processing a 1D signal. Furthermore, the convolutional kernels in original CNN are randomly initialized and there is no pretraining process. In this paper, we constructed a 1D CNN that is suitable for processing the 1D primary spatiotemporal feature. In addition, we also incorporate convolutional autoencoder (CAE) and linear autoencoder (AE) to initialize the CNN. Through this pretraining process, the performance of SACONVA is increased.

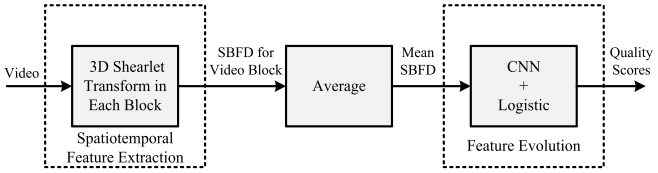


Fig. 1. Overview framework of SACONVA.

The remainder of the paper is organized as follows. Section II introduces the detailed structure and related techniques about the proposed framework. In Section III, experimental results and a thorough analysis of this framework are presented. In Section IV, a real application of SACONVA is presented. Finally, the conclusion and future works are given in Section V.

II. Methodology

The proposed framework of using 3D shearlet transform and CNN for video quality estimation is shown in Fig. 1. The main problems that need to be considered in this framework include: 1) spatiotemporal feature extraction using 3D shearlet transform and 2) feature evolution using CNN. More details about these problems will be described in the following sections.

A. 3D Shearlet Transform

It is known that traditional wavelets and their associated transforms are highly efficient when approximating and analyzing 1D signals. However, these frameworks have some limitations when extended to process multidimensional data such as images or videos. Typically, multidimensional data exhibit curvilinear singularities and wavelets cannot effectively detect their directions and in the sense sparsely approximate them. To overcome the drawbacks of wavelets, a new class of multiscale analysis methods has been proposed in recent years, which is defined as the third-generation wavelet. A noteworthy characteristic of these new methods is their ability to efficiently capture anisotropic features in multidimensional data and the shearlet representation is one of them. 3D shearlet transform [19], [20] is a natural extension of the original 2D shearlet transform [21]–[27]. Similar to the 2D version, to achieve the multiscale and multidirectional analysis, the scaling, shearing, and translating matrixes are well defined in three pyramidal regions and the general case of the 3D universal shearlet system is defined as the collections

$$\begin{aligned} &\mathrm{SH}(\phi, \psi^{(1)}, \psi^{(2)}, \psi^{(3)}; a, k, c) \\ &\quad = \Phi(\phi; c_1) \cup \Psi^{(1)}(\psi^{(1)}; a, k, c) \cup \Psi^{(2)}(\psi^{(2)}; a, k, c) \\ &\qquad \cup \Psi^{(3)}(\psi^{(3)}; a, k, c) \end{aligned} \tag{1}$$

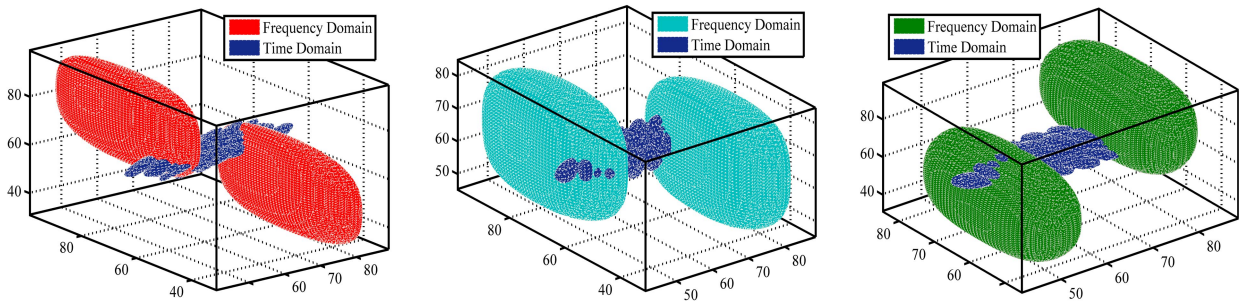


Fig. 2. 3D shearlets in the time domain and the frequency domain.

where

$$\begin{cases} \Phi(\phi; c_1) = \{\phi_m = \phi(\cdot - c_1 m) : m \in Z^2\} \\ \Psi^{(d)}(\psi^{(d)}; \alpha, k, c) \\ = \left\{ \psi^{(d)}_{j,k,m} = 2^{\frac{a_j+1}{4} j} \psi^{(d)} \left(S^{(d)}_k A^{(d)}_{a_j, 2^j} \cdot - M^{(d)}_c m \right) : \right. \\ \left. \quad j \geq 0, \quad |k| \leq \lceil 2^{j(a_j-1)/2} \rceil, \quad m \in Z^2 \right\} \end{cases} \tag{2}$$

where $c = (c_1, c_2) \in (\mathbf{R}_+)^2$ and $k = (k_1, k_2) \in \mathbf{Z}^2$. The scaling matrix $A^{(d)}_{a,2^j}$, the shearing matrix $S^{(d)}_k$, and the translating matrix $M^{(d)}_c$ are defined by

$$A^{(1)}_{a,2^j} = \begin{pmatrix} 2^j & 0 & 0 \\ 0 & 2^{aj/2} & 0 \\ 0 & 0 & 2^{aj/2} \end{pmatrix}, \quad A^{(2)}_{a,2^j} = \begin{pmatrix} 2^{aj/2} & 0 & 0 \\ 0 & 2^j & 0 \\ 0 & 0 & 2^{aj/2} \end{pmatrix}$$

$$A^{(3)}_{a,2^j} = \begin{pmatrix} 2^{aj/2} & 0 & 0 \\ 0 & 2^{aj/2} & 0 \\ 0 & 0 & 2^j \end{pmatrix}$$

$$S^{(1)}_k = \begin{pmatrix} 1 & k_1 & k_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad S^{(2)}_k = \begin{pmatrix} 1 & 0 & 0 \\ k_1 & 1 & k_2 \\ 0 & 0 & 1 \end{pmatrix}$$

$$S^{(3)}_k = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ k_1 & k_2 & 1 \end{pmatrix}$$

$$M^{(1)}_c = \mathrm{diag}(c_1, c_2, c_2), \quad M^{(2)}_c = \mathrm{diag}(c_2, c_1, c_2),$$
$$M^{(3)}_c = \mathrm{diag}(c_2, c_2, c_1)$$

where $a = (a_j)_j \in (0, 2)$, $a_j \in (0, 2)$, and $d = 1, 2, 3$. Fig. 2 shows the 3D shearlets both in the time and in the frequency domain. Similar to 2D shearlets, in the time domain, the 3D shearlets are rotatable wavelets that are compactly supported. In the frequency domain, they are well-defined directional filter banks that tile the entire frequency space. The ability of the 3D shearlet transform to deal with geometric information efficiently and its sparsity properties have the potential to produce a significant improvement in NR-VQA application.

B. Feature Extraction

Natural videos possess substantial spatiotemporal correlations. They do not change randomly over space or time; instead, video frames at different times and spatial positions are highly correlated. This property is usually referred as NSS and a lot of works have been done to propose NSS models for natural images. However, there are few NSS models for natural videos. In [16], [28], and [29], the statistical properties of the frame-differenced natural videos are explored in DCT, fast Fourier transform, and wavelet domains and shown that the frame-differenced natural videos reliably obey a space-time spectral model. When the distortions are introduced into natural videos, the property of the distorted video will become different from the natural video, which causes deviations from NSS models and these deviations can be applied as the indicator of video quality. The NSS property of frame-differenced natural videos can be easily represented using the mean shearlet-based feature descriptors (SBFDs) of each video block. We start the derivation of SBFD for one video block and the calculation process is summarized in Fig. 3. Each element x_i is defined as

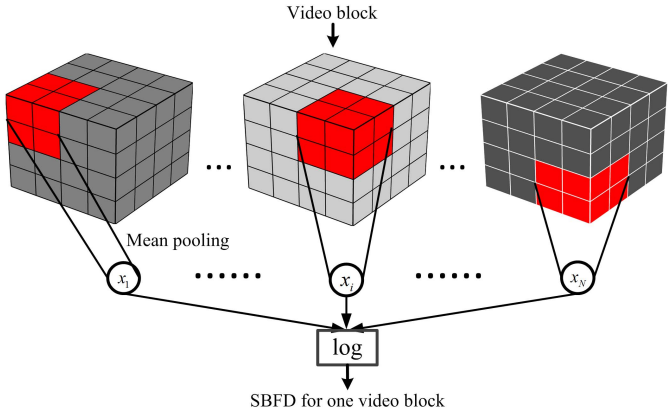


Fig. 3. Calculation process of SBFD for one video block. The large cubes with different colors stand for shearlet coefficients in different sub-bands. The small red blocks stand for pooling regions.

$$x(a, k, b) = \frac{\sum_{c \in b} |\mathrm{SH}_\phi v(a, k, c)|}{\prod_{i=1}^{3} m_i} \tag{3}$$

where $a = 1, \ldots, A$ is the scale index (exclude coarsest scale), $k = 1, \ldots, K$ is the direction index, $b = 1, \ldots, \prod_{i=1}^{3} (M_i / m_i)$ is the index of pooling regions (the red blocks in Fig. 3) in each sub-band, and c is the time shift. M represents the size of the video block (the same as shearlet coefficients) and m indicates the size of the pooling region. $\mathrm{SH}_\phi v(a, k, c)$ is the shearlet coefficient at the particular scale, direction, and time. One thing to note here is that the definition of a, k, and c in (1) is a little different from (3). In (1), they stand for variables. In (3), they represent index.

After the mean pooling of shearlet coefficients in each pooling region, the pooled values are concatenated as a vector and every element in this vector is subject to a logarithmic

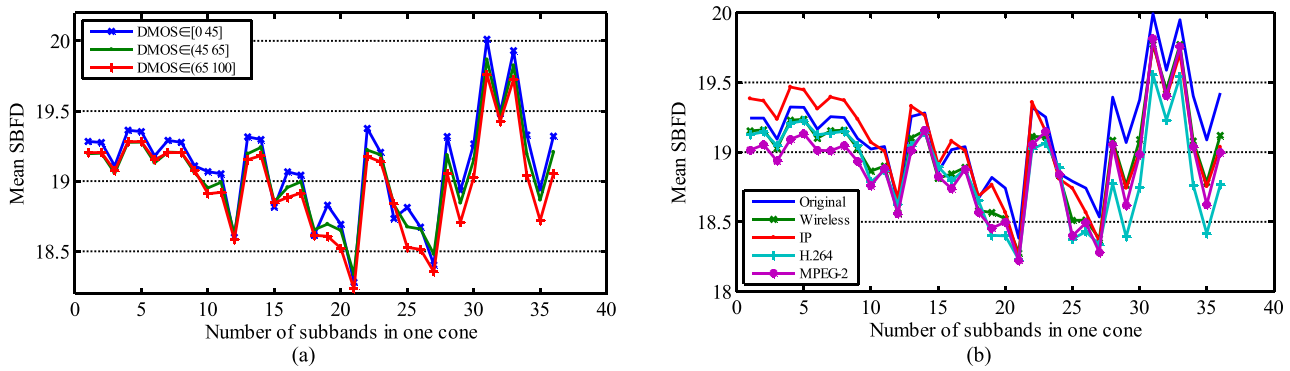


Fig. 4. (a) Mean SBFD versus sub-band enumeration index for video blocks with different DMOSs in LIVE VQA database. (b) Mean SBFD versus sub-band enumeration index for natural video blocks and different distorted video blocks in LIVE VQA database. To make the figure clear, only the SBFD features in one pyramidal region are shown. Original: the original reference video. Wireless: wireless distortions. IP: IP distortions. H.264: H.264 compression. MPEG-2: MPEG-2 compression.

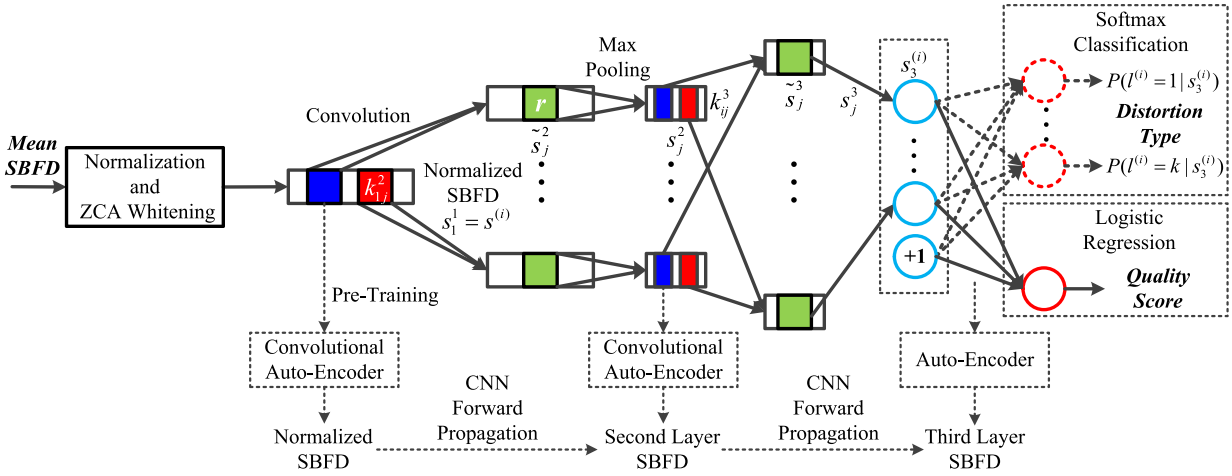


Fig. 5. Detailed architecture of SACONVA.

nonlinearity. The SBFD for each video block is represented as

$$\text{SBFD} = (\log(x_1), \log(x_2), \ldots, \log(x_N)) \tag{4}$$

where $N = A \times K \times \prod_{i=1}^{3} (M_i/m_i)$ is the total number of pooling regions and is also the length of the SBFD. The final SBFD of the whole video is the average of the SBFD for each video block.

We segment all the videos of the laboratory for image & video engineering (LIVE) VQA database into $128 \times 128 \times 128$ (horizontal $\times$ vertical $\times$ temporal) video block, extract SBFD in each block, and calculate the mean SBFD for video blocks with different differential MOSs (DMOSs) and with different distortion types. The pooling region we adopted here is the same size as the video block. To generate Fig. 4(a), we divide all the video blocks into three groups based on their DMOS and calculate the mean SBFD of video blocks in each group. We can see that the mean SBFD decreases as the amount of perceived distortion in the video increases. Fig. 4(b) plots the mean SBFD versus sub-band enumeration index for original video blocks and four different distorted video blocks. It can be clear seen that the SBFD differentiates the original video blocks from different distorted video blocks. Saad *et al.* [16] observed that the distribution of the original video DCT coefficients is more heavy tailed than that of the distorted video DCT coefficients, which means that more video DCT coefficients become smaller after distortion. This property is also confirmed by SBFD.

C. Feature Evolution and Quality Estimation

Krizhevsky *et al.* [30], Zeiler and Fergus [31], Jain and Seung [32], Xie *et al.* [33], Larochelle *et al.* [34], Erhan *et al.* [35], Goodfellow *et al.* [36], and Hinton and Salakhutdinov [37] demonstrated and provided a clear understanding of why deep neural networks perform so well in image classification work, and our previous works have also discovered that stacked AEs can help to exaggerate the discriminative parts of primary features and made them more distinguishable. In this paper, we propose applying the CNN as the feature evolution process and the primary SBFDs are evolved by this network before being sent to the softmax classifier. The proposed 1D CNN consists of five layers. Two fully connected convolutional layers, two max pooling layers and one output layer, which are shown in Fig. 5. The primary advantage of using multiple layer CNN is that it will allow us to compute much more complex features of the input signal. Because each hidden layer computes a nonlinear transformation of the previous layer, a deep network can have a significantly greater representational power than does a shallow one. Before being sent into the CNN, the input SBFD is normalized by subtracting the mean and dividing by the standard deviation of its elements, and zero components

analysis whitening is performed to the normalized SBFD. Suppose we have a training set $\{(s^{(1)}, l^{(1)}), \ldots, (s^{(P)}, l^{(P)})\}$ of P labeled data, where the input is $s^{(i)}$, which is the normalized mean SBFD for a video and the length is N. $l^{(i)}$ is the label of this video, which is normalized DMOS in quality estimation task or distortion type in distortion classification task.

The activations of the convolution layer can be computed as

$$\tilde{s}_j^{\ell} = f\left(\sum_{i \in Q_j} s_i^{\ell-1} * k_{ij}^{\ell} + b_j^{\ell}\right) \tag{5}$$

where k_{ij}^{ℓ} is the kernel weights, b_j^{ℓ} is the bias, and Q_j identifies the group of latent feature maps and the convolution is the valid border handling type. The activation function $f(x)$ is set to the sigmoid function in this paper.

In the CNN, after the convolutional layer, a pooling layer is usually added to compute a lower resolution and translation-invariant representation of the convolutional layer activations through subsampling. Max-pooling is a form of nonlinear down sampling. There are two reasons why we are using max-pooling. The first reason is to eliminate nonmaximal values and reduce computation for upper layers. The second reason is that it provides a form of translation invariance and provides additional robustness to position. Therefore, max-pooling is a good way of reducing the dimensionality of intermediate representations. The max-pooling activation can be computed as

$$s_j^{\ell}(m) = \max_{k=1}^{r} \left(\tilde{s}_j^{\ell}((m-1) \times r + k)\right) \tag{6}$$

where s_j^{ℓ} is the pooling layer's output of the jth feature map. r is the pooling size that indicates the number of samples to be pooled together.

In the derivation that follows, we will consider the squared-error loss function. The physical meaning of cost function is to make sure the output of the neural network is exactly the same as the true label when training. Based on different tasks, the loss function is different. For quality estimation task, the output layer is a logistic regression and the error is defined as

$$J(\theta) = \frac{1}{2P} \sum_{i=1}^{P} \left\| s_3^{(i)} - l^{(i)} \right\|^2 \tag{7}$$

where $s_3^{(i)}$ is the final output of CNN and $l^{(i)}$ is the corresponding video label. For distortion classification task, however, the output layer is a softmax classifier and the error is given by

$$J(\theta) = -\frac{1}{P}\left[\sum_{i=1}^{P}\sum_{j=0}^{1} 1\{l^{(i)} = j\} \log p\left(l^{(i)} = j | s_3^{(i)}; \theta\right)\right] \tag{8}$$

and

$$p\left(l^{(i)} = j | s_3^{(i)}; \theta\right) = \frac{e^{\theta_j^T s_3^{(i)}}}{\sum_{l=1}^{K} e^{\theta_l^T s_3^{(i)}}} \tag{9}$$

where $1\{\cdot\}$ is the indicator function, which means $1\{\text{a true statement}\} = 1$, and $1\{\text{a false statement}\} = 0$. K is the number of distortion type and θ is the softmax parameter vector.

In the traditional CNN, convolutional kernels and the weights in output layer are usually randomly initialized. In this paper, we initialize the convolutional kernels using several CAEs [38] with the same topology and initialize the output layer weights using the linear AE. Through this unsupervised pretraining process, the network can alleviate common problems with training deep neural networks, such as converging to local optima and diffusion of gradients.

The kernels in each convolutional layer can be initialized by the CAE with the same topology as the CNN. The CAE architecture is intuitively the combination of weights share mechanism employed in the CNN and the general idea of AE. Similar to the CNN, the latent representation of the jth feature map for a monochannel input s_i^{l-1} is defined as

$$h_j = f\left(s_i^{l-1} * k_j + b_j\right). \tag{10}$$

Since the CAE has only one convolutional layer and the input is only an SBFD vector in our method, (10) is actually the same as the first convolutional layer in (5) when $i = 1$. Further, max-pooling layer is introduced to h_j by setting all nonmaximal values to zero in nonoverlapping subregions, which is the same as expanding the down-sampled data after max-pooling in CNN backpropagation process. We define the output of max-pooling layer as h_j, which is the same length as h_j. The reconstruction is obtained using

$$t^{(i)} = f\left(\sum_{j \in H} h_j * k_j + c\right). \tag{11}$$

In (11), a single bias c per latent map is used. H is the group of latent feature maps and k_j is the flipped version of k_j. The cost function of the CAE is also the same as that of (7), where $s_3^{(i)}$ is substituted by $t^{(i)}$. One thing should be noted is that since the normalized SBFD is not in the range of 0–1, the sigmoid function is not used when pretraining the first convolution layer.

The weights in the final output layer are initialized by the linear AE. The forward propagation of the linear AE is defined as

$$\begin{cases} z^{(2)} = \mathrm{W}^{(1)} s_3^{(i)} + b^{(1)} \\ a^{(2)} = f\left(z^{(2)}\right) \\ \hat{s}_3^{(i)} = \mathrm{W}^{(2)} a^{(2)} + b^{(2)}. \end{cases} \tag{12}$$

Then we define the overall cost function to be

$$J(W,b) = \left[\frac{1}{2P}\sum_{i=1}^{P}\left(\left\|\hat{s}_3^{(i)} - s_3^{(i)}\right\|\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{2}\sum_{i=1}^{q_l}\sum_{j=1}^{q_{l+1}}\left(W_{ji}^{l}\right)^2. \tag{13}$$

The linear AE has parameters $(W,b) = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)})$, where $W_{ji}^{(l)}$ indicates the weight associated with the connection between unit i in layer l and unit j in layer $l+1$. q_l is the number of neurons in layer l and λ is called weight decay parameter, which controls the relative importance of the two terms. The second weight decay term is used to decrease the magnitude of the weights and helps prevent overfitting.

If we further incorporate sparsity parameter ρ, the overall cost function is modified as

$$J_{\text{sparse}}(W,b) = J(W,b) + \beta \sum_{j=1}^{q_2} \text{KL}(\rho||\rho j) \tag{14}$$

where ρ controls the weight of the sparsity penalty term. After using this constraint, the hidden unit's activations must mostly be near 0. KL is the Kullback–Leibler divergence defined as

$$\text{KL}(\rho||\hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j} \tag{15}$$

and

$$\hat{\rho}_j = \frac{1}{2P} \sum_{i=0}^{P} \left[a_j^{(2)}(s_3^i) \right]. \tag{16}$$

In sum, we can apply very similar architecture to deal with both video quality estimation and video distortion classification task.

III. Experiments and Related Analysis

A. Experimental Protocol

1) Data Set: To effectively evaluate the performance of SACONVA and other video quality assessment algorithms, the following three publicly available video-quality databases are used, which contain multiple types of distortion.

1) *The LIVE Video Quality Database [39]:* The LIVE VQA database contains 10 reference videos and 150 distorted videos. There are four types of distortion, which include wireless distortions, IP distortions, H.264 compression, and MPEG-2 compression. All video files are in raw YUV 4:2:0 format and the spatial resolution of all videos is 768 × 432 pixels. DMOS and the standard deviation of the DMOS are provided for each video.
2) *The image & video processing laboratory (IVPL) Video Quality Database [40]:* The IVPL HD VQA database consists of 10 reference videos and 128 distorted videos. There are four types of distortion in this database, which include Dirac wavelet compression, H.264 compression, simulated transmission of H.264-compressed bitstreams through error-prone IP networks, and MPEG-2 compression. All video files are in raw 4:2:0 format and the spatial resolution of all videos is 1920 × 1088 pixels. DMOS and the standard deviation of the DMOS are also provided for each video. The DMOSs in this database are derived from nonexperts, experts, and all observers, respectively. In this paper, we use the DMOS derived from experts.
3) *The CSIQ Video Quality Database [41]:* The CSIQ VQA database consists of 12 reference videos and 216 distorted videos. All videos in this database are in raw YUV 4:2:0 format and the spatial resolution of all videos is 832 × 480 pixels. Each reference video has 18 distorted versions with six types, which include motion JPEG (MJPEG), H.264, HEVC, wavelet compression using SNOW codec, packet-loss in a simulated wireless network, and additive white Gaussian noise (noise). DMOS and the standard deviation of the DMOS for each video are also given.

2) Parameters of SACONVA: 3D shearlet transform is applied to each $128 \times 128 \times 128$ video block and each frame of the tested video is converted into grayscale image. The video block is decomposed into four scales (exclude approximation component) and the direction number for each scale is nine. However, in the real implementation of the 3D filter banks, some regions are overlapped among three cones. Therefore, we have not adopted the full shearlet system and omitted certain shearlets lying on the borders of the second and third pyramids. The pooling region we adopted in the experiments is the same size as the video block, and the final primary SBFD feature for each video is a 52 vector. The results are expected to be further improved if we reduce the size of pooling region, but the cost is larger feature size. For example, if we adopt 64 × 64 × 64 as the pooling region size, the median distortion classification accuracy of LIVE database will improve from 0.7750 to 0.8478, but the feature length is increased from 52 to 416. For CNN, the kernel number and kernel size are 10 and 19, respectively, for the first convolution layer, and 100 and 10, respectively, for the second convolution layer. The max-pooling size is two for each subsampling layer. In addition, we use the limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm to optimize the cost function of the CNN and CAE. The key parameter for L-BFGS is the maximum number of iterations and it is set as 400 when training CNN and CAE (the reason why we set it as 400 is explained in the Appendix).

3) Evaluation: Two measures are chosen to evaluate the performance of VQA algorithms: 1) linear correlation coefficient (LCC) and 2) Spearman rank order correlation coefficient (SROCC). In each train-test iteration, we randomly select 80% of reference videos and their distorted versions as the training set and the remaining 20% as the test set.

B. Discussion About SACONVA

1) Effects of Feature Evolution Process: In Section II-C, we have mentioned that the CNN can be used to exaggerate the discriminative parts of the primary SBFD and made them more distinguishable. In this section, we will demonstrate that the quality estimation performance will be significantly improved after bringing in the feature evolution process. To demonstrate the effect of feature evolution process, the primary SBFD is first sent into the logistic regression directly, and then CNN with one and two convolution layers are added between the SBFD and logistic regression. We have compared the quality estimation performance of these three situations. Fig. 6 shows the boxplots of comparison results, which are obtained from 100 train-test iterations using LIVE database. It can be clearly seen that the performance is increased if the primary SBFD is evolved before sending into the logistic regression.

2) Effects of CNN Pretraining: It is also mentioned in Section II-C that the quality estimation performance can be further improved if the convolution kernels are pretrained using a CAE instead of random initialization. In this section, we also provide experiments to demonstrate the effect of CNN pretraining. To simplify the experimental process, the CNN we adopted in this experiment only contains one convolution layer. Fig. 7 shows how the quality estimation

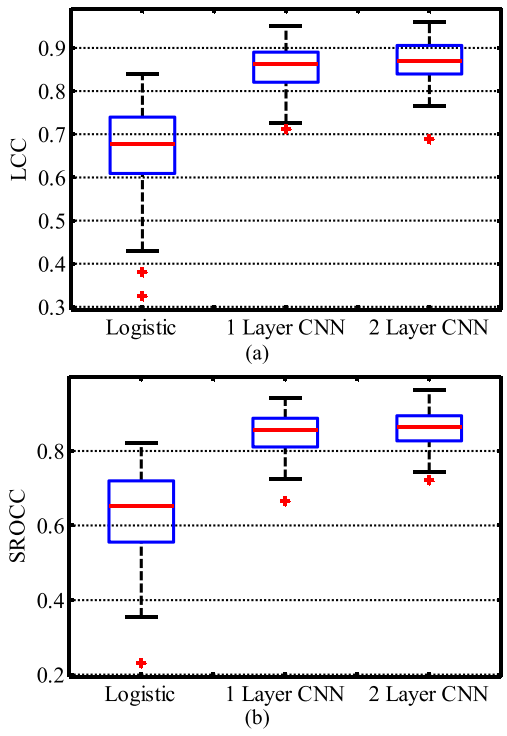


Fig. 6. Box plot of LCC and SROCC distributions versus number of convolution layers from 100 runs of experiments on the LIVE database. Logistic means the primary SBFD features are directly sent into the logistic regression. (a) Box plot of LCC distributions. (b) Box plot of SROCC distributions.

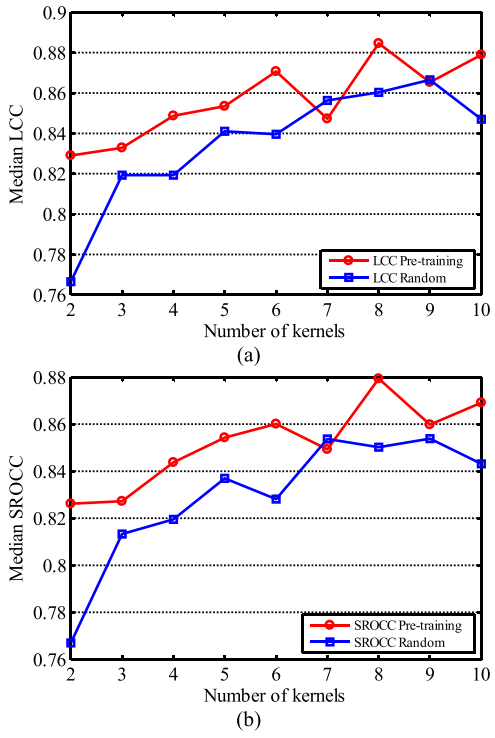


Fig. 7. Plot of median LCC and SROCC versus number of convolution kernels from 100 runs of experiments on the LIVE database. (a) Plot of median LCC. (b) Plot of median SROCC.

performance varies with the number of convolution kernels both for CAE pretraining and random initialization. It is not surprising to find that: 1) the performance exhibits the rising tendency with the increase in kernel numbers and

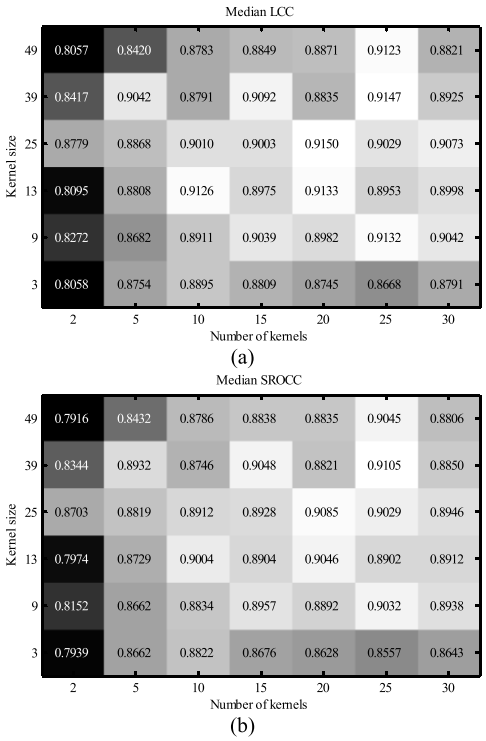


Fig. 8. 2D plot of median LCC and SROCC for different combinations of kernel number and kernel size from 100 runs of experiments on the LIVE database. (a) 2D plot of median LCC. (b) 2D plot of median SROCC.

2) the performance of pretraining is almost always better than random initialization.

3) Effects of CNN Parameters: Several parameters, such as number of kernels and kernel size, are involved in the design of CNN. In this section, we examine how these parameters affect the performance of SACONVA on the LIVE database. We train and test the CNN using different combinations of kernel number and kernel size and show the median LCC and SROCC for each combination. Fig. 8 shows how the performance changes with the kernel number and kernel size. We can see that the performance shows the rising tendency with the increase in kernel number, not being very sensitive to kernel size.

C. Performance Evaluation

In this section, we test SACONVA on LIVE, IVPL, and CSIQ databases, and compare it with the state-of-the-art FR and NR approaches. Four FR-I/VQA methods include: peak-signal-to-noise ratio (PSNR), structural similarity index (SSIM), VIF, ST-MAD, and ViS3. For PSNR, SSIM, and VIF, we obtain them for each individual video frame and calculate the average of every frame as evaluation for the whole video. One general purpose NR-VQA methods is video BLIINDS. The source code of PSNR, SSIM, and VIF we used is available at (http://foulard.ece.cornell.edu/gaubatz/metrix_mux/#modernization). The source code for STMAD and ViS3 we employed is available at (http://vision.okstate.edu/?loc=home), and the source code of video BLIINDS can be obtained at (http://live.ece.utexas.edu/research/quality/). The output layer of SACONVA is logistic regression, which can be seen as a

TABLE I

MEDIAN LCC AND SROCC CORRELATIONS FOR 100 ITERATIONS OF EXPERIMENTS ON THE LIVE VQA DATABASE. (ALGORITHMS ARE NR-VQA ALGORITHMS)

	LCC					SROCC				
	Wireless	IP	H.264	MPEG-2	ALL	Wireless	IP	H.264	MPEG-2	ALL
PSNR	0.7277	0.6383	0.7359	0.6535	0.7499	0.7381	0.6000	0.7143	0.6327	0.6958
SSIM	0.7969	0.8269	0.7110	0.7849	0.7883	0.7381	0.7714	0.6905	0.7846	0.7211
VIF	0.7468	0.6916	0.6954	0.7505	0.7541	0.7143	0.6000	0.5476	0.7319	0.6873
STMAD	0.8887	0.8956	0.9209	0.8992	***0.8777***	0.8257	0.7714	0.9323	0.8733	0.8301
ViS3	0.8597	0.8576	0.7809	0.7650	0.8251	0.8257	0.7714	0.7657	0.7962	0.8156
V-BLIINDS	***0.9357***	***0.9291***	0.9032	0.8757	0.8433	0.8462	0.7829	0.8590	***0.9371***	0.8323
SACONVA(S)	0.9019	0.9271	***0.9336***	***0.9176***	0.8714	***0.8810***	***0.8286***	***0.9471***	0.9018	***0.8569***
SACONVA(N)	0.8455	0.8280	0.9116	0.8778		0.8504	0.8018	0.9168	0.8614	

TABLE II

MEDIAN LCC AND SROCC CORRELATIONS FOR 100 ITERATIONS OF EXPERIMENTS ON THE IVPL VQA DATABASE. (ALGORITHMS ARE NR-VQA ALGORITHMS)

	LCC					SROCC				
	MPEG-2	IP	H.264	Dirac	ALL	MPEG-2	IP	H.264	Dirac	ALL
PSNR	0.7851	0.8394	0.8818	0.8806	0.8054	0.6571	0.8286	0.8095	0.8286	0.7887
SSIM	0.7352	0.6218	0.7488	0.8036	0.6605	0.6000	0.7629	0.6667	0.6571	0.6459
VIF	0.9205	0.6879	0.8095	0.8890	0.7264	0.8286	0.7671	0.7545	0.8767	0.6892
STMAD	0.9162	0.8451	***0.9427***	0.9399	0.8809	0.8286	0.7943	***0.9048***	0.8932	***0.8862***
ViS3	0.9021	0.7968	0.8927	0.8794	0.8313	0.8407	0.8114	0.8095	0.8857	0.8528
V-BLIINDS	0.9186	0.8963	0.8876	0.9385	0.8476	***0.8918***	0.8286	0.8321	***0.8976***	0.8315
SACONVA(S)	***0.9346***	***0.9134***	0.9203	***0.9521***	***0.8860***	0.8688	***0.8578***	0.8981	0.8809	0.8695
SACONVA(N)	0.8811	0.7499	0.8529	0.9165		0.8286	0.7514	0.8389	0.8772	

TABLE III

MEDIAN LCC AND SROCC CORRELATIONS FOR 100 ITERATIONS OF EXPERIMENTS ON THE CSIQ VQA DATABASE. (ALGORITHMS ARE NR-VQA ALGORITHMS)

	LCC							**SROCC**						
	H.264	**PLoss**	**MJPEG**	**Wavelet**	**Noise**	**HEVC**	**ALL**	**H.264**	**PLoss**	**MJPEG**	**Wavelet**	**Noise**	**HEVC**	**ALL**
PSNR	0.9208	0.8246	0.6705	0.9235	0.9321	0.9237	0.7137	0.8810	0.7857	0.6190	0.8810	0.8333	0.8571	0.7040
SSIM	0.9527	0.8471	0.8047	0.8907	***0.9748***	0.9652	0.7627	***0.9286***	0.8333	0.6905	0.8095	***0.9286***	***0.9148***	0.7616
VIF	0.9505	***0.9212***	***0.9114***	0.9241	0.9604	0.9624	0.7282	0.9048	***0.8571***	0.8095	0.8571	0.8810	0.9012	0.7256
STMAD	***0.9619***	0.8793	0.8957	0.8765	0.8931	0.9274	0.8254	0.9286	0.8333	0.8333	0.8095	0.8095	0.8810	0.8221
ViS3	0.9356	0.8299	0.8110	***0.9303***	0.9373	***0.9677***	0.8100	0.9286	0.8095	0.7857	***0.9048***	0.8571	0.9025	0.8028
V-BLIINDS	0.9413	0.7681	0.8536	0.9039	0.9318	0.9214	0.8494	0.9048	0.7481	0.8333	0.8571	0.9048	0.8810	0.8586
SACONVA(S)	0.9392	0.8180	0.8686	0.9103	0.9516	0.9146	***0.8668***	0.9086	0.8031	***0.8350***	0.8571	0.9024	0.8574	***0.8637***
SACONVA(N)	0.9133	0.8115	0.8565	0.8529	0.9027	0.9068		0.9048	0.7840	0.7857	0.8333	0.8810	0.8333	

special case of generalized linear model and thus analogous to linear regression. Since the output of the logistic regression is in the range of 0–1, we also normalized the DMOS of each database into the same range when training the CNN. The normalization is given by

$$\hat{x} = \frac{x - \min(x)}{\max(x - \min(x))} \tag{17}$$

where x indicates a vector that contains all the DMOS of each database. $\hat{x}$ is the normalized DMOS vector and every element in this vector is in the range of 0–1. Before computing LCC and SROCC, the predicted DMOS can be mapped back to the original range. (Actually, since this normalization is a linear map, it has no effect on the results of LCC and SROCC no matter we map the predicted DMOS back to the original range or not.) In addition, to make a fair comparison, we perform a nonlinear mapping on the predicted scores produced by FR-I/VQA methods to transform the quality measure into the same range with DMOS. We apply a four-parameter logistic transform to the raw predicted scores, as recommended by video quality experts group in [42]. The four-parameter logistic transform is given by

$$f(x) = \frac{\tau_1 - \tau_2}{1 + \exp\left(-\frac{x - \tau_3}{|\tau_4|}\right)} + \tau_2 \tag{18}$$

where x denotes the raw predicted score and τ_1, τ_2, τ_3, and τ_4 are free parameters that are selected to provide the best fit of the predicted scores to the subjective rating scores. Just as the same procedure with training and testing NR-VQA algorithms, we split each VQA database into two parts of 80% and 20%. Each time, 80% of data is used for estimating parameters of the logistic function and 20% is used for testing. All results reported in this section are obtained by 100 train-test iterations. The algorithms were separately tested on those portions of each database that contain specific distortions as well as on the entire database containing all the distortions mixed together. Besides, for SACONVA, we conducted both distortion-specific experiments and nondistortion-specific experiments. In distortion-specific experiments, we train and test on each of the distortions in the databases. In nondistortion-specific experiments, we separate the relative test distortion set and use all the other videos in each database as training set.

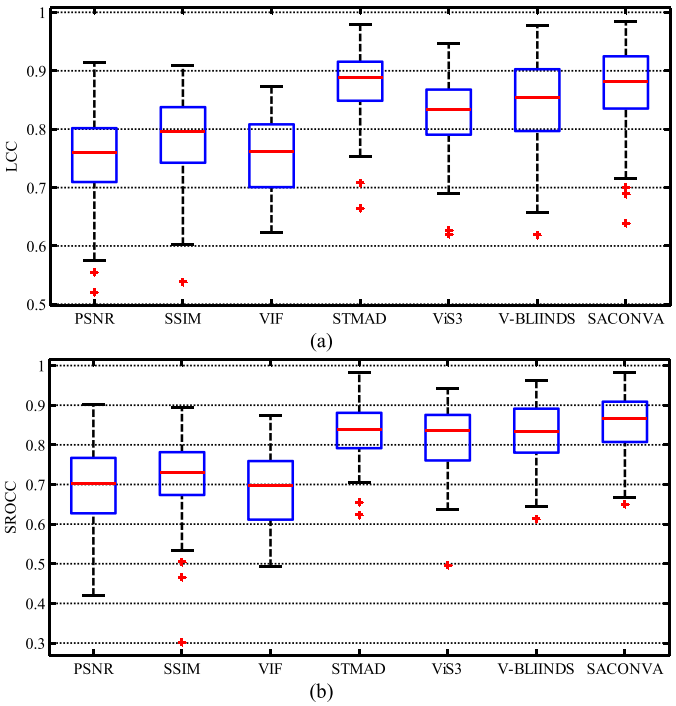


Fig. 9. Box plot of LCC and SROCC distributions of the algorithms over 100 trails on the LIVE VQA database. (a) Box plot of LCC distribution. (b) Box plot of SROCC distribution.

The testing results of the three databases are listed in Tables I–III, respectively. SACONVA(S) indicates the distortion-specific experiments and SACONVA(N) denotes the nondistortion-specific experiments. It can be seen that the performance of SACONVA outperforms the state-of-the-art NR-VQA method V-BLIINDS and is close to the FR-I/VQA methods. To visualize the statistical significance of the comparison, we show the box plot of LCC and SROCC distributions of different I/VQA approaches from 100 runs of experiments on the LIVE VQA database in Fig. 9(a) and (b). It is clear that SACONVA performs well among all the measures under consideration. We also show the scatter plots of the predicted DMOS versus DMOS on the test sets for the entire three VQA databases in Fig. 10(a)–(c). These scatter plots exhibit nice linear relationship along each axis and we also list the LCC and SROCC of this run for reference. Besides, we visualize the learned convolution kernels on LIVE VQA database in Fig. 11, which represent the discriminative features of SBFD generated by supervised learning. We may assume that these features are just like basis functions with different scales, and the input SBFD is expanded on these functions. Furthermore, to observe how the number of training set affects the overall performance, we vary the percentage of training and testing set and plot the median performance for LIVE database. Fig. 12 shows the change of median performance with respect to the training percentage. We can observe that a larger number of training data leads to higher performance.

D. Distortion Classification

As shown in Fig. 5, SACONVA is not only applicable to video quality estimation but also can be easily extended to video distortion identification. To demonstrate this, we still use LIVE, IVPL, and CSIQ VQA databases to test the classification ability of SACONVA. Similar to quality estimation, in this paper, SACONVA is also trained using mean SBFD and the distortion label for each video. We report the median classification accuracy of the softmax classifier for each of the distortions and all distortions in the three VQA databases in Table IV. To visualize the classification performance, we also plot the mean confusion matrix for each distortion type and the box plot for all distortions across 100 train-test trials, which are shown in Figs. 13 and 14, respectively. We can see from the testing results that the distortion types in CSIQ database are much easier to be identified. Wireless and IP distortion are most confused with each other in LIVE database. MPEG-2 and Dirac are most confused with each other in IVPL database. However, H.264 distortion is generally slightly confused with other distortions in all the three databases. Besides, the reason why the accuracy of the original video is low is that there exists training bias in all the three databases. The number of the distorted video for each distortion is three or four times larger than that of the original video. Thus, the classification accuracy of original video can be further improved if more original video samples are incorporated.

E. Computational Cost

Our experiments are performed on a PC with 3.40-GHz CPU and all the approaches are implemented using MATLAB script and tested under MATLAB R2013a environment. We randomly select 10 videos from LIVE database and calculate the median processing time for each method. There are two main steps involved in SACONVA. The feature evolution step is not very time consuming. The training time of CNN for the entire LIVE database is 13.3064 s and the testing time is only 0.005835 s. Intuitively, the SBFD extraction step spends much time. However, this step is highly parallelizable. First, the SBFD for each video block can be extracted in parallel. Second, in the implementation of 3D shearlet, the multiplication between frequency domain and filter banks is also parallel. In addition, a further computational advantage can be attained since the CNN training process and the filter bank calculation process can be completed before the real prediction.

To provide a fair comparison, we first implemented SACONVA using sequential execution. Then, we optimized it using a multithreading technique that executes SACONVA in parallel. Fig. 15 plots the median computation time for each method. It can be clear seen that the execution time can be reduced if we consider the highly parallelizable property of SACONVA.

IV. Application to Blind Video Denoising

Video denoising is a fundamental and desirable work in many applications, such as video enhancement, compression, and pattern recognition. Usually, the original videos are needed to obtain the optimized parameters for video denoising algorithms. However, in many scenarios, it is very hard or expensive to get them. Therefore, in this section, we propose to use SACONVA for accomplishing the blind video denoising work. SACONVA can automatically determine the threshold

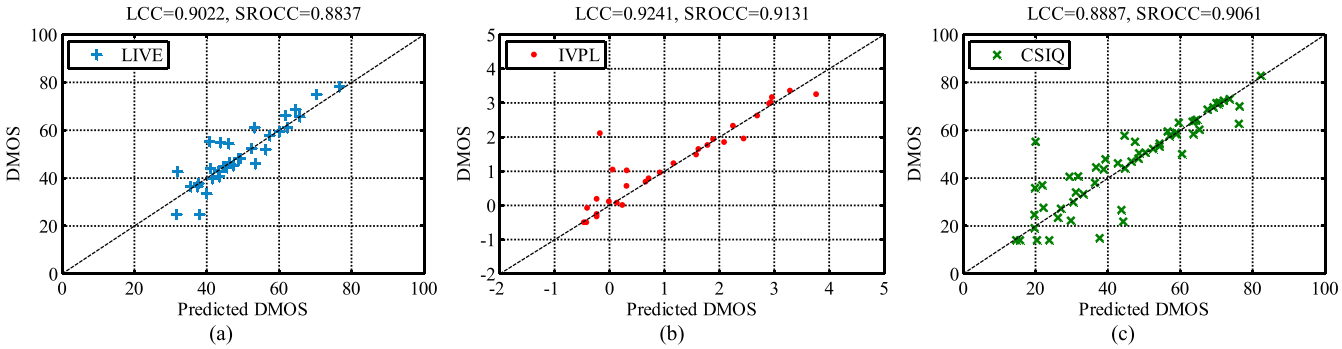


Fig. 10. Predicted DMOS versus subjective DMOS on the three databases. (a) Scatter plot on LIVE. (b) Scatter plot on IVPL. (c) Scatter plot on CSIQ.

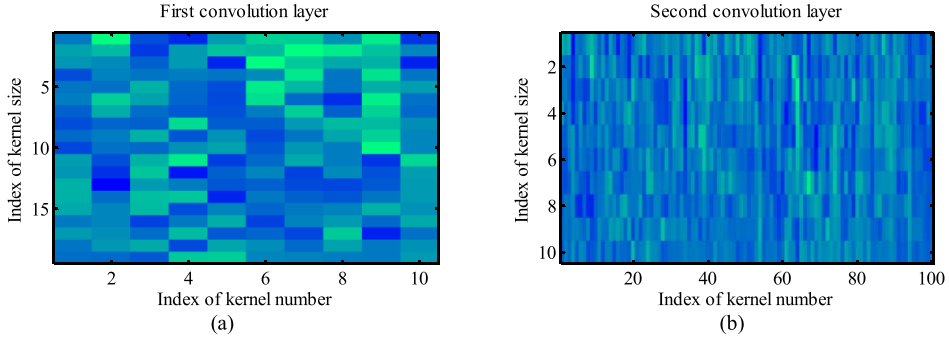


Fig. 11. Learned convolution kernels on LIVE VQA database. (a) First convolution layer kernels. (b) Second convolution layer kernels.

TABLE IV
MEDIAN CLASSIFICATION ACCURACY FOR 100 ITERATIONS OF EXPERIMENTS ON THREE DATABASES

	LIVE							
	Ori	**Wireless**	**IP**	**H.264**	**MPEG-2**	**ALL**		
Accuracy	0.6392	0.7575	0.5972	0.8148	0.9028	0.7750		
	IVPL							
	Ori	**MPEG-2**	**IP**	**H.264**	**Dirac**	**ALL**		
Accuracy	0.5367	0.7884	0.7247	0.8897	0.8926	0.8116		
	CSIQ							
	Ori	**H.264**	**PLoss**	**MJPEG**	**Wavelet**	**Noise**	**HEVC**	**ALL**
Accuracy	0.5125	0.8985	0.8854	0.8737	0.9329	0.9373	0.9052	0.8859

of shearlet denoising algorithm and guides this algorithm to recover the original video from Gaussian noise.

One sample 3D shearlet video denoising algorithm for additive Gaussian noise is hard threshold method, which is defined as

$$\mathrm{SH}_\phi v(a,k,c) = \begin{cases} \mathrm{SH}_\phi v(a,k,c), & \text{if } |\mathrm{SH}_\phi v(a,k,c)| \\ & \quad > T \times \mathrm{RMS}(a,k) \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

where T is the threshold and root mean squares (RMS) is the RMS of all shearlets and can be calculated by

$$\mathrm{RMS}(a,k) = \sqrt{\frac{\sum_c |\mathrm{SH}_\phi v(a,k,c)|^2}{\prod_{i=1}^{3} M_i}}. \tag{20}$$

RMS can be used to normalize shearlet coefficients to make them comparable and are determined if the filter banks are specified. Thus, in this method, T is important to the final result. In this application, we assume the standard variance σ of Gaussian noise is unknown. To get the value of T, we can

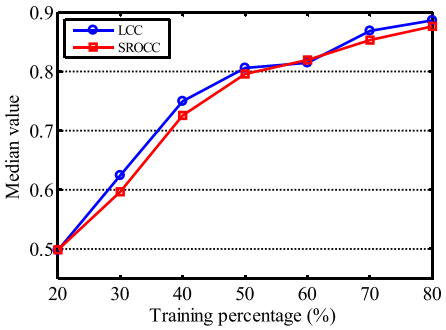


Fig. 12. Median LCC and SROCC with respect to the training percentage over 100 runs on LIVE database.

consider an objective function, which is defined as

$$T^* = \arg\max_T \mathrm{QualityCost}(T). \tag{21}$$

The objective function QualityCost means the quality between denoised video and original video. Through optimizing this function, the optimized threshold T^* can be obtained. If the original video is known, the objective function can be

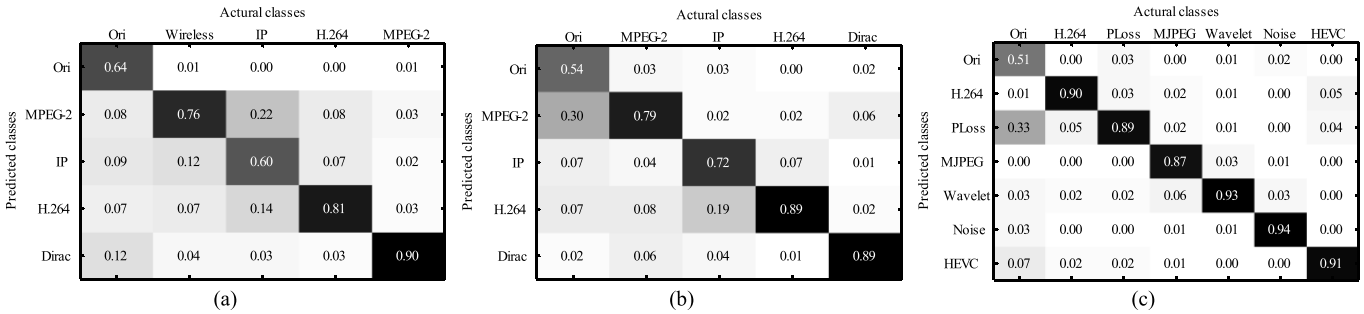


Fig. 13. Mean confusion matrix for the softmax classifier across 100 trials. (a) Mean confusion matrix of LIVE VQA database. (b) Mean confusion matrix of IVPL VQA database. (c) Mean confusion matrix of CSIQ VQA database.

TABLE V
SUMMARY OF 3D SHEARLET DENOISING ALGORITHM

Data: v_{ori}, v_{noi} **Result**: v^*_{PSNR}, $v^*_{SACONVA}$ $T := T_{init}$ $T^*_{PSNR} := \underset{T}{argmax}\ \mathrm{PSNRCost}(T, v_{ori}, v_{noi})$ $T^*_{SACONVA} := \underset{T}{argmax}\ \mathrm{SACONVACost}(T, v_{noi})$ $v^*_{PSNR} := S^{-1}T^*_{PSNR}S(v_{noi})$ $v^*_{SACONVA} := S^{-1}T^*_{SACONVA}S(v_{noi})$	$Q := \mathrm{PSNRCost}(T, v_{ori}, v_{noi})$ $v_{PSNR} := S^{-1}TS(v_{noi})$ $Q := \mathrm{PSNR}(v_{ori}, v_{PSNR})$	$Q := \mathrm{SACONVACost}(T, v_{noi})$ $v_{SACONVA} := S^{-1}TS(v_{noi})$ $Q := \mathrm{SACONVA}(v_{SACONVA})$
S represents the 3D shearlet transform, S^{-1} represents the inverse 3D shearlet transform and T means hard threshold denoising with threshold T. $S^{-1}TS(v)$ indicates a process of shearlet transform, hard threshold denoising and inverse shearlet transform of video v.		

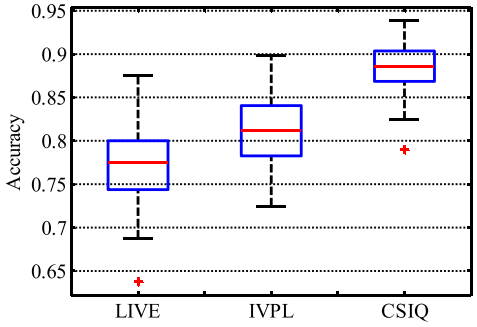


Fig. 14. Box plot of classification accuracy on three databases over 100 trails.

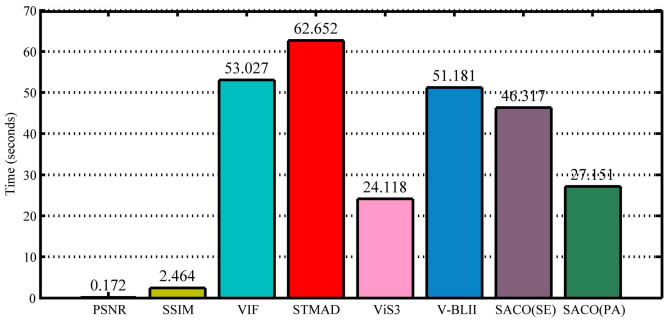


Fig. 15. Median computation time for each VQA methods.

substituted by any FR-VQA method, in this application, we use PSNR as the FR-VQA and (21) can be expressed as

$$T^*_{\mathrm{PSNR}} = \arg\max_{T} \mathrm{PSNRCost}(T, v_{\mathrm{ori}}, v_{\mathrm{noi}}) \tag{22}$$

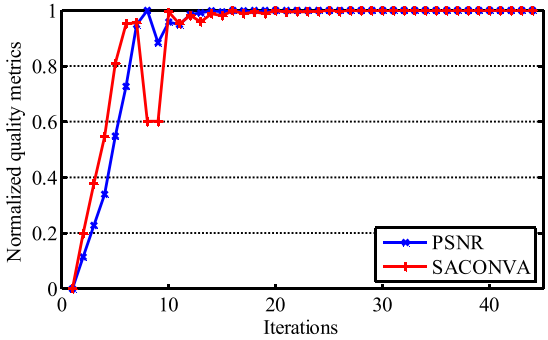


Fig. 16. Normalized cost function values versus iterations.

where v_{ori} and v_{noi} indicate the original video and noisy video, respectively. However, if the original video cannot be obtained, we use SACONVA as objective function and (21) can be expressed as

$$T^*_{\mathrm{SACONVA}} = \arg\max_{T} \mathrm{SACONVACost}(T, v_{\mathrm{noi}}). \tag{23}$$

The detailed implementation of this algorithm is summarized in Table V. We trained SACONVA using the entire CSIQ VQA database since this database contains Gaussian noise distortion type, and tested the algorithm on the standard *Coastguard* video sequence. Gaussian noise with three different standard variances σ (30, 40, and 50) is added to the original video sequence. Fig. 16 shows the convergence process of the optimization algorithm. To compare this

TABLE VI
VIDEO DENOISING PERFORMANCE USING DIFFERENT COST FUNCTIONS

		Standard variance		
		20	**30**	**40**
Threshold	PSNRCost	49.0029	76.4058	105.7598
	SACONVACost	54.1932	81.3286	112.5410
PSNR	Noisy	22.1092	18.5923	16.0910
	PSNRCost	26.7930	24.6442	23.3976
	SACONVACost	26.1324	24.5415	23.2823

process, we normalize the output values [using (17)] of two cost functions and plot them in one coordinate. Table VI lists the video denoising performance and optimized thresholds using different cost functions. It can be seen that PSNR shows good performance, since it has original video as reference. However, without the reference image, the performance of SACONVA is still good. It also guides the denoising algorithm to find a reasonable threshold and helps to improve the final PSNR between denoised video and original video.

V. CONCLUSION

In this paper, we have proposed a general-purpose NR-VQA algorithm SACONVA, which is developed based on the 3D shearlet transform and CNN. We have extracted simple and efficient primary spatiotemporal features SBFD using a 3D shearlet transform and evolved SBFD using CNN to make them more discriminative. The final score is given by a simple logistic regression. SACONVA is tested on LIVE, IVPL, and CSIQ VQA databases and compared with state-of-the-art FR-I/VQA and NR-VQA approaches. SACONVA correlates highly with human perception and is highly comparative with the state-of-the-art VQA methods. In addition, we also conducted several experiments to demonstrate that SACONVA can be easily extended to complete distortion identification work and blind video denoising work.

APPENDIX

The maximum number of iterations is a key parameter for L-BFGS when training the CNN and CAE. The underfitting or overfitting problem will be caused if it is not properly tuned. In the experiments, we have found that the number of iterations for training the CNN is much more important than the CAE, and the underfitting or overfitting problem of the whole framework is usually caused by it when training the CNN. Therefore, in this Appendix, we focus on discussing how to select a proper iteration number when training the CNN.

To select a proper parameter, we have plotted the median LCC and SROCC for both training set and testing set in LIVE database when using different iteration numbers. It can be observed from Figs. 17 and 18 that with the increasing of iteration number, the performance of the training set increases monotonously. However, the performance of the testing set increases first and then decreases, which means the testing set performance is a convex function of iteration number and it goes through from underfitting stage to best fitting stage and then to overfitting stage. This convex function achieves its maximum value when the iteration number is around 400.

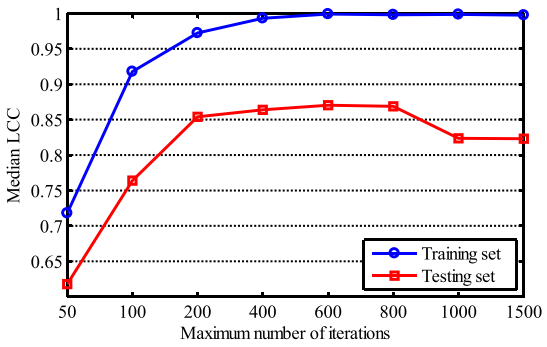


Fig. 17. Median LCC of the training and testing sets versus maximum number of iterations over 100 runs on LIVE database.

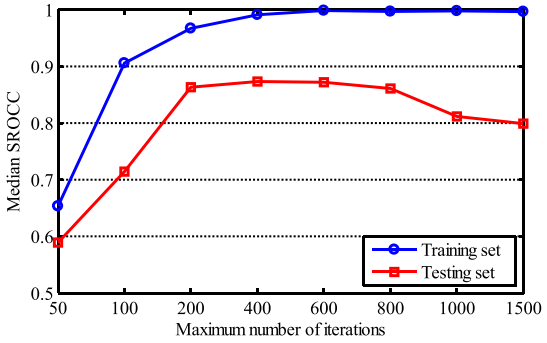


Fig. 18. Median SROCC of the training and testing set versus maximum number of iterations over 100 runs on LIVE database.

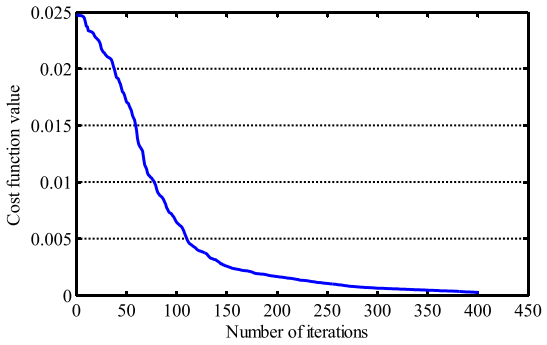


Fig. 19. Cost function value of CNN versus iterations.

Therefore, we set the maximum number of iterations as 400 (for LIVE database and other databases). Besides, it can be also observed from Fig. 19 that after 400 iterations, the cost function of CNN is already tend to converge.

REFERENCES

[1] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[2] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *Proc. 1st Int. Workshop Video Process. Quality Metrics Consum. Electron.*, 2005, pp. 23–25.

[3] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[4] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *Proc. 18th IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2011, pp. 2505–2508.

[5] P. V. Vu and D. M. Chandler, "ViS$_3$: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *J. Electron. Imag.*, vol. 23, no. 1, pp. 013016-1–013016-25, 2014.

[6] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[7] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[8] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. Int. Conf. Image Process.*, vol. 1. 2002, pp. I-477–I-480.

[9] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1918–1927, Nov. 2005.

[10] T. Brandão and M. P. Queluz, "No-reference quality assessment of H.264/AVC encoded video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1437–1447, Nov. 2010.

[11] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

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

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

[14] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1098–1105.

[15] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1733–1740.

[16] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, Mar. 2014.

[17] J. Xu, P. Ye, Y. Liu, and D. Doermann, "No-reference video quality assessment via feature learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 491–495.

[18] Y. Li, L.-M. Po, X. Xu, and L. Feng, "No-reference image quality assessment using statistical characterization in the shearlet domain," *Signal Process., Image Commun.*, vol. 29, no. 7, pp. 748–759, 2014.

[19] P. S. Negi and D. Labate, "3-D discrete shearlet transform and video processing," *IEEE Trans. Image Process.*, vol. 21, no. 6, pp. 2944–2954, Jun. 2012.

[20] G. Kutyniok, W.-Q. Lim, and R. Reisenhofer. (2014). "Shearlab 3D: Faithful digital shearlet transforms based on compactly supported shearlets." [Online]. Available: http://arxiv.org/abs/1402.5670

[21] S. Yi, D. Labate, G. R. Easley, and H. Krim, "A shearlet approach to edge analysis and detection," *IEEE Trans. Image Process.*, vol. 18, no. 5, pp. 929–941, May 2009.

[22] G. Easley, D. Labate, and W.-Q. Lim, "Sparse directional image representations using the discrete shearlet transform," *Appl. Comput. Harmon. Anal.*, vol. 25, no. 1, pp. 25–46, 2008.

[23] G. Kutyniok and W.-Q. Lim, "Image separation using wavelets and shearlets," in *Curves and Surfaces*. Berlin, Germany: Springer-Verlag, 2011.

[24] G. Kutyniok, W.-Q. Lim, and X. Zhuang, "Digital shearlet transforms," in *Shearlets*. Boston, MA, USA: Birkhäuser, 2012, pp. 239–282.

[25] G. Kutyniok, M. Shahram, and X. Zhuang, "ShearLab: A rational design of a digital parabolic scaling algorithm," *SIAM J. Imag. Sci.*, vol. 5, no. 4, pp. 1291–1332, 2011.

[26] D. L. Donoho, G. Kutyniok, M. Shahram, and X. Zhuang, "A rational design of a digital shearlet transform," *Proc. 9th Int. Conf. Sampling Theory Appl.*, Singapore, 2011.

[27] G. Kutyniok, M. Shahram, and D. L. Donoho, "Development of a digital shearlet transform based on pseudo-polar FFT," *Proc. SPIE*, vol. 7446, p. 74460B, Sep. 2009.

[28] D. W. Dong and J. J. Atick, "Statistics of natural time-varying images," *Netw., Comput. Neural Syst.*, vol. 6, no. 3, pp. 345–358, 1995.

[29] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2013.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[31] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2013, pp. 818–833.

[32] V. Jain and S. Seung, "Natural image denoising with convolutional networks," in *Proc. NIPS*, vol. 8. 2008, pp. 769–776.

[33] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Proc. NIPS*, 2012, pp. 350–358.

[34] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An empirical evaluation of deep architectures on problems with many factors of variation," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 473–480.

[35] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *J. Mach. Learn. Res.*, vol. 11, pp. 625–660, Feb. 2010.

[36] I. J. Goodfellow, Q. V. Le, A. M. Saxe, H. Lee, and A. Y. Ng, "Measuring invariances in deep networks," in *Proc. NIPS*, vol. 9. 2009, pp. 646–654.

[37] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[38] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Artificial Neural Networks and Machine Learning*. Berlin, Germany: Springer-Verlag, 2011, pp. 52–59.

[39] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[40] Image & Video Processing Laboratory, The Chinese University of Hong Kong. (Apr. 20, 2012). *IVP Subjective Quality Video Database*. [Online]. Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective/index.shtml

[41] Laboratory of Computational Perception & Image Quality, Oklahoma State University. (Nov. 15, 2012). *CSIQ Video Database*. [Online]. Available: http://vision.okstate.edu/?loc=stmad, accessed 2013.

[42] Video Quality Expert Group, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II," Video Quality Expert Group, Boulder, CO, USA, Tech. Rep., 2003.

Yuming Li (S'13) received the B.E. and M.E. degrees from Huazhong University of Science and Technology, Wuhan, China, in 2011 and 2013, respectively. He is working toward pursuing the Ph.D. degree with City University of Hong Kong, 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 City University of Hong Kong, Hong Kong, in 1988 and 1991, respectively.

He has been with the Department of Electronic Engineering, City University of Hong Kong, since 1991, where he is currently an Associate Professor and a Laboratory Director of TI Educational Training Centre. He has authored over 140 technical journal and conference papers. His research interests include image and video coding with an emphasis on fast encoding algorithms, new motion compensated prediction techniques, and 3D video processing.

Dr. Po is a member of the Technical Committee on Multimedia Systems and Applications and the IEEE Circuits and Systems Society. He was the Chairman of the IEEE Signal Processing Hong Kong Chapter in 2012 and 2013. He was an Associate Editor of *HKIE Transactions* in 2011 to 2013. He also served on the Organizing Committee of the IEEE International Conference on Acoustics, Speech and Signal Processing in 2003, and the IEEE International Conference on Image Processing in 2010.

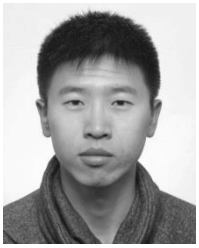
Chun-Ho Cheung received the B.Eng. (Hons.) degree in computer engineering and the Ph.D. degree in electronic engineering from City University of Hong Kong, Hong Kong, in 1996 and 2002, respectively.

He is an Assistant Professor with the Department of Information Systems, City University of Hong Kong. His research interests include image coding, motion estimation, and e-Learning.

Xuyuan Xu (S'11) received the B.E. degree in information engineering from City University of Hong Kong, in 2010, where he is currently working toward the Ph.D. degree with the Department of Electronic Engineering. His B.E. thesis was entitled Stereoscopic Video Generation from Monoscopic Video.

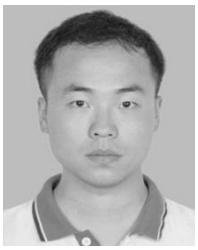
His research interests include 3D video coding and 3D view synthesis.

Mr. Xu won the Best Tertiary Student Project of the Asia Pacific International and Communication Award in 2010.



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 the Tianjin Jinhang Institute of Technical Physics, Tianjin, China, in 2011. He is currently working toward the Ph.D. degree with 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 the 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 with the Department of Electronic Engineering, City University of Hong Kong.

His research interests include image and biomedical signal processing, and machine learning.

Kwok-Wai Cheung received the B.E., M.S., and Ph.D. degrees from City University of Hong Kong, Hong Kong, in 1990, 1994, and 2001, respectively, all in electronic engineering.

He was a Research Student/Research Assistant with the Department of Electronic Engineering, City University of Hong Kong, from 1996 to 2002. He joined Chu Hai College of Higher Education, Hong Kong, in 2002, where he is currently an Associate Professor with the Department of Computer Science. His research interests include image/video coding and multimedia database.