DEEP LEARNING NETWORK FOR BLIND IMAGE QUALITY ASSESSMENT

Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang

Insti. of Image Commu. & Infor. Proce., Shanghai Jiao Tong Univ., Shanghai, China, 200240
Shanghai Key Laboratory of Digital Media Processing and Transmissions
E-mail: gukesjtuee@sjtu.edu.cn

ABSTRACT

Nowadays, blind image quality assessment (BIQA) has been intensively studied with machine learning, such as support vector machine (SVM) and k-means. Existing BIQA metrics, however, do not perform robust for various kinds of distortion types. We believe this problem is because those frequently used traditional machine learning techniques exploit shallow architectures, which only contain one single layer of nonlinear feature transformation, and thus cannot highly mimic the mechanism of human visual perception to image quality. The recent advance of deep neural network (DNN) can help to solve this problem, since the DNN is found to better capture the essential attributes of images. We in this paper therefore introduce a new Deep learning based Image Quality Index (DIQI) for blind quality assessment. Extensive studies are conducted on the new TID2013 database and confirm the effectiveness of our DIQI relative to classical full-reference and state-of-the-art reduced- and no-reference IQA approaches.

Index Terms— Image quality assessment (IQA), blind / no-reference (NR), machine learning, deep learning

1. INTRODUCTION

Human society is ushering into the information age of high-speed development. Due to the constant proliferation of high-volume visual data, the task to faithfully assess image quality is becoming increasingly important, and this renders the related image quality assessment (IQA) much more attractive than before. Furthermore, IQA can help for compression [1]-[3], interpolation [4]-[5], and denoising [6], [21]. We can roughly divide IQA metrics into two classes: subjective assessment and objective assessment. The former one is usually considered to be the terminal quality gauge, since it applies a series of complicated, expensive, laborious and time-consuming steps to derive the raw scores, i.e. mean opinion score (MOS), from inexperienced subjects. Present popular and large image quality databases mainly include LIVE [7], TID2008 [8], CSIQ [9], and TID2013 [10]. But it is easy to view that subjective IQA methods cannot be widely employed, particularly in real-time applications, and this results in a growing number of objective IQA metrics in recent years.

Objective assessment methods can be further categorized into three types: 1) full-reference (FR) [11]-[15]; 2) reduced-reference (RR) [16]-[18]; 3) no-reference (NR) or blind image quality assessment (BIQA) [19]-[25]. The last type of BIQA algorithms is a currently hot topic, and has been deeply explored during the last three years. Quite a few BIQA methods (e.g. DIIVINE [19], BLIINDS-II [20], and BRISQUE [21]) were designed to first extract features from distorted images based on the natural scene statistics (NSS) [26], and then to learn a regression module trained from feature space to subjective scores. Very lately, researchers began to focus on developing NR IQA metrics without human scored images, prior knowledge of image scenes and distortion categories, for instance, natural image quality evaluator (NIQE) [24] and quality-aware clustering (QAC) [25].

Existing BIQA models mainly resort to machine learning techniques, such as the support vector machine (SVM) [27] which is used in [19]-[23], and k-means [28] which is used in [25]. Those NR IQA approaches are found to work effectively for commonly encountered compression artifacts, Gaussian blur and white noise. But unfortunately, they are not available for other more kinds of distortion types, which is probably due to the fact that those used traditional machine learning methods exploit shallow architectures with a single layer of nonlinear feature transformation. As a result, the mechanism of human visual perception to image quality cannot be well approximated. The recently developed deep neural network (DNN) [29] is shown to capture images' fundamental properties very well. This provides a new solution to deal with this problem. We accordingly propose a new Deep learning based Image Quality Index (DIQI) for blind quality assessment. Experimental results suggests that the DNN is a highly promising tool in developing IQA tasks in comparison to traditional machine learning techniques.

The remainder of this paper is organized as follows. Section 2 first reviews the deep learning model. Section 3 then describes the implementation of our DIQI in detail. In Section 4, a comparison of the DIQI and classical FR, as well as state-of-the-art RR and NR IQA algorithms is conducted on the newly released TID2013 database. Section 5 finally concludes this paper.

2. DEEP LEARNING MODEL

2.1. Neural network

In most cases, we adopt linear and logistic regressions to do simple regression or classification tasks. But they cannot work fine when those models are used to solve a problem with many input features. To that end, some improved models like nonlinear regression have been designed. However, we still always trap into the trouble of choosing high order features, and furthermore, those models are easy to cause the over fitting problem. Neural network is found to be a good simulation of human brain's neurons and their connections. Every neuron is expressed by a simple logistic regression, and some neurons together form one layer of neural network. Since each layer gives a nonlinear output, this model can solve nonlinear problems without worry about how to choose nonlinear features. Scientists have found out that multi-layer neural network has the better ability to learn features more intrinsic than input ones. But we usually encounter three problems: 1) multi-layer neural network is hard to train, because it will take a lot of time to converge the randomly initialized weights of each neuron for the whole network; 2) too many layers are easy to generate gradient diffusion problems; 3) the neural network easily hovers near the local minimum. Hence, the DNN was put forward recently.

2.2. Deep learning

Deep learning is a multi-layer-structure neural network with an input layer, multiple hidden layers and an output layer. Hinton came up with this in 2006, in order to overcome problems of traditional neural networks [29]. In the DNN network, only the adjacent layers are connected, and neurons in every hidden layer are logistic regression models. The output layer will be logistic regression or softmax regression for classification tasks, while be linear regression for regression tasks, as illustrated in Fig. 1. The idea of DNN is realized as follows: 1) to train the parameters of each hidden layer using unlabeled data, making each layer learn the structure of original data and represent it more abstractly; 2) to fine tune the pre-learned parameters with labeled data until the final converge to the global optimum.

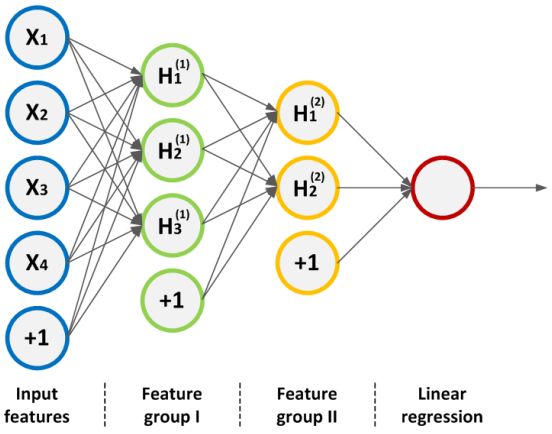


Fig. 1. Deep neural network.

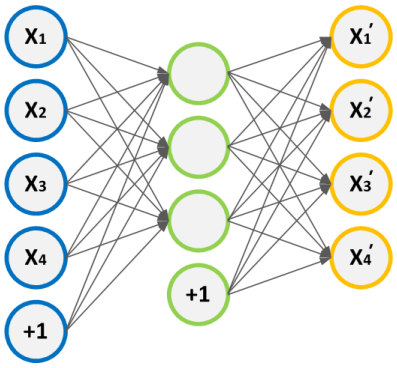
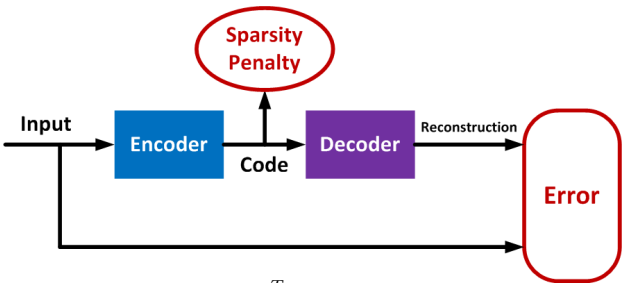


Fig. 2. Sparse autoencoder.

2.3. Sparse autoencoder

The most important part of deep learning model is to pre-train each layer, which leads to the remarkable difference between traditional neural networks and DNN network. In this research, the sparse autoencoder is adopted to pre-train each layer. Sparse autoencoder has three layers, and the input and output layers have the same number of units, as shown in Fig. 2. Let the input features be $\{X_1, X_2, X_3, X_4\}$. We hope to obtain the same output $\{X_1, X_2, X_3, X_4\}$. That is to say, this model represents the information translation process without any loss. But this layer has only three neurons, and it cannot gives a completely abstract description of the original data. As a consequence, we just want the output to be $\{X_1', X_2', X_3', X_4'\}$ with the loss information as little as possible. Furthermore, note that neurons in human brain are not always active. They behave active at some time while inactive at most of other time. We thus add L_1 regularity in this model to regulate only a few neurons to be non-zero yet most of neurons to be zero in each layer. Fig. 3 exhibits a primary framework of sparse representation.



- Input: X - Code: $h = W^T X$

- Loss: $L(X, W) = \| Wh - X \|^2 + \lambda \sum_j |h_j|$

Fig. 3. Sparse representation.

3. DEEP LEARNING BASED IMAGE QUALITY INDEX

In [21] and [24], Mittal *et al*. utilized the classical NSS model in the spatial domain [26], which can validly capture the essential low-order statistics of natural images. For an input image I, the spatial NSS model works with the process of local mean removal and divisive normalization:

$$\tilde{I}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + 1} \tag{1}$$

with

$$\mu(i,j) = \sum_{k=-K}^{K} \sum_{l=-L}^{L} w(k,l) \cdot I(i+k, j+l) \tag{2}$$

$$\sigma(i,j) = \sqrt{\sum_{k=-K}^{K} \sum_{l=-L}^{L} w(k,l) \cdot [I(i+k,j+l) - \mu(i,j)]^2} \tag{3}$$

where $\mathbf{w} = \{w(k,l) | k = -K, ..., K; l = -L, ..., L\}$ indicates a 2D circularly-symmetric Gaussian weighting function, and K and L are assigned as 3.

The normalized coefficients in Eq. (1) were observed to reliably follow a Gaussian distribution when computed from natural images without apparent distortion [26]. This ideal model, however, is violated when the input images suffer distortions. The degree of modification can measure perceptual distortion severity. For instance, those coefficients of natural images exhibits a Gaussian-like appearance, while the Gaussian blur makes natural images a more Laplacian appearance, as shown in Fig. 4. In addition, the moment-matching based generalized Gaussian distribution (GGD) is found to effectively catch a broader spectrum of statistics of distorted images [30]. This parametric model is thereby deployed to fit distributions of those coefficients from original and distorted images. For each input image, we from the GGD fit of those coefficients estimate two parameters (γ, σ^2), which separately control the distribution's 'shape' and 'variance'.

Despite its successfulness, we found three problems in the above-mentioned solution: 1) the distribution of those normalized coefficients cannot completely towards standard distributions, e.g. Gaussian distribution; 2) the parameters of the GDD fit are just approximately estimated with the method in [30]; 3) those two estimated parameters hardly characterize the input image well. An easy way to solve these problems is to put the whole distribution of $\tilde{I}$ as input features when training. But the input features seems extremely large. It is very lucky that the DNN network is very suitable for this. The concrete steps used in this work are as follows:

• First, to extract features from an input image. We convert the input RGB image to YIQ color space, before estimating the normalized coefficients from the luminance information Y and the chrominance information I and Q. We then calculate the distribution of Y, I and Q, extracting a total number of 3000 features from an image.

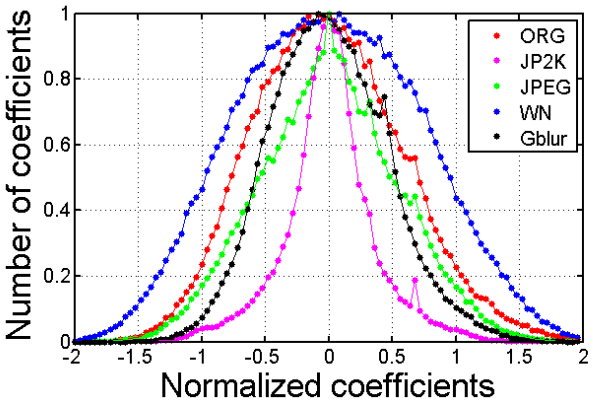


Fig. 4. Histogram of normalized coefficients for a natural undistorted image and its various distorted versions from the TID2013 database [10]. ORG: the original image. JP2K: JPEG2000 compression. JPEG: JPEG compression. WN: additive white Gaussian noise. Gblur: Gaussian blur.

• Second, to initialize the DNN network. The DNN is set with three hidden layers. We choose the 'minFunc' tools based on the L-BFGS algorithm [31] to train the first sparse autoencoder. Input data is a matrix of size $\mathbf{s} \times 3000$, including $\mathbf{s}$ training samples with 3000 features in each sample. We first compute activation of hidden layer using sigmoid function, and then compute output using linear function. The compute cost is

$$J_{sparse}(W,b) = J(W,b) + \beta \sum_{j=1}^{s_2} KL(\rho \parallel \hat{\rho}_j) \tag{4}$$

where

$$J(W,b) = \left[\frac{1}{m}\sum_{j=1}^{m} J(W,b;x^{(i)},y^{(i)})\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(W_{ji}^{(l)})^2$$

$$= \left[\frac{1}{m}\sum_{j=1}^{m}(\frac{1}{2}\parallel h_{W,b}(x^{(i)}) - y^{(i)} \parallel^2)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(W_{ji}^{(l)})^2$$

and

$$KL(\rho \parallel \hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j}$$

with ρ and $\hat{\rho}$ being the mean activation and expected mean activation. We use the back propagation algorithm to compute the gradient for each layer. The maximum iteration of L-BFGS is set to be 1000. On this basis, we implement this cost function to train the first sparse autoencoder. Thereafter, we train the second and third sparse autoencoders using the similar way like the first sparse autoencoder. Notice that the input data is the output of the last sparse autoencoder. We finally train the linear regression layer with the output of sparse autoencoders. Although the output layer is linear regression and definitely our data does not distribute linearly, the two hidden

layer use nonlinear functions and the whole model is still nonlinear regression. The cost function is

$$J(W,b) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(Y - Label)^2 + \frac{\lambda}{2}\sum W_i^2. \quad (5)$$

• Third, to fine tune the DNN network. This model is a neural network model except that all the weights are initialized using the pre-trained weights. We use the back propagation algorithm to tune the weights in each layer, with the maximum iteration of L-BFGS being 400.

4. EXPERIMENTAL RESULTS

The newly released TID2013 database [10] was employ in this paper. TID2013 is composed of 3000 distorted images generated by corrupting 25 source images (24 natural images and one artificial image) with 24 distortion types at 5 distortion levels. These distortion types include: 1) additive white Gaussian noise; 2) additive white Gaussian noise which is more intensive in color components than in the luminance component; 3) additive Gaussian spatially correlated noise; 4) masked noise; 5) high frequency noise; 6) impulse noise; 7) quantization noise; 8) Gaussian blur; 9) image denoising; 10) JPEG compression; 11) JPEG2000 compression; 12) JPEG transmission errors; 13) JPEG2000 transmission errors 14) non-eccentricity pattern noise; 15) local block-wise distortion of different intensity; 16) mean shift; 17) contrast change; 18) change of color saturation; 19) multiplicative Gaussian noise; 20) comfort noise; 21) lossy compression of noisy images; 22) image color quantization with dither; 23) chromatic aberrations; 24) sparse sampling and reconstruction. It is noted that the proposed DIQI metric depends on the statistics of natural images, so we only selected those 24 natural images and their associated distorted versions, summing to 2880 images. We then measure the performance indices of DIQI and nine IQA metrics, which include: 1) classical FR PSNR, SSIM [11], IFC [12], VIF [13]; 2) recently proposed RR FEDM [16], SDM [17]; 3) state-of-the-art NR / blind BRISQUE [21], NIQE [24], QAC [25].

To account for the correlation performance of the DIQI, a training procedure is required to calibrate the regressor module. Similar to the usual training method, we randomly separate 2880 distorted images in TID2013 into two subsets. One is the training set which consists of distorted images corresponding to 75% original images, and the other is the testing set containing the rest 25% distorted images. In order to ensure that the DIQI is robust across image contents and is not governed by the specific train-test split, this paper repeats this random 75% train - 25% test procedure 1000 times, and report the median result of the performance across these 1000 iterations so as to eliminate performance bias as much as possible. The commonly used performance evaluation, Spearman rank-order correlation coefficient (SROCC) [32], is applied to measure those competing IQA algorithms. We report the performance indices in Table 1. It can be easily find that the proposed DIQI metric achieves inspiring results. It is comparable to those testing FR IQA methods, and superior to recently developed RR and NR IQA algorithms.

Table 1. Comparison of performance evaluations of the proposed DIQI and nine competing IQA algorithms.

Metrics	Type	SROCC	Metrics	Type	SROCC
PSNR	FR	0.6393	SDM	RR	0.4670
SSIM	FR	0.6273	BRISQUE	NR	0.5258
IFC	FR	0.5389	NIQE	NR	0.3115
VIF	FR	0.6768	QAC	NR	0.3721
FEDM	RR	0.3061	DIQI	NR	0.6728

5. CONCLUSION

In this paper, we investigate the problem of deep learning network for blind image quality assessment. Most existing NR IQA metrics were proposed with the great support of machine learning techniques, e.g. SVM, and have acquired fairly well performance. Those BIQA algorithms, however, are available for some commonly encountered distortion types, whereas they work ineffectively for other types of image distortions. This problem may be explained by the fact that those frequently used machine learning methods are built upon shallow architectures and cannot well approximate the sensation of HVS to image quality. We use the recently proposed deep learning model to solve this problem, since it exploits the multi-layer structure and is easy to establish and train. Taking 3000 numbers as input features for one image and the deep learning model as machine learning tool, this paper proposes a new Deep learning based Image Quality Index (DIQI) for blind quality assessment. Experimental results on the newly released TID2013 database are provided to confirm the effectiveness of the proposed DIQI algorithm against nine classical FR IQA approaches and state-of-the-art RR and NR IQA metrics, and the DNN is a promising tool in the IQA research.

Acknowledgment

This work was supported in part by NSFC (61025005, 61371146, 61221001), 973 Program (2010CB731401) and FANEDD (201339).

6. REFERENCES

[1] G. Zhai, J. Cai, W. Lin, X. Yang, and W. Zhang, "Three dimensional scalable video adaptation via user-end perceptual quality assessment," *IEEE Trans. Broadcasting*, vol. 54, no. 3, pp. 719-727, September 2008.

[2] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, "Cross-dimensional perceptual quality assessment for low bitrate videos," *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1316-1324, November 2008.

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

[4] X. Liu, D. Zhao, R. Xiong, S. Ma, W. Gao, and H. Sun, "Image Interpolation via Regularized Local Linear Regression," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3455-3469, December 2011.

[5] X. Liu, D. Zhao, J. Zhou, W. Gao, and H. Sun, "Image Interpolation via Graph-based Bayesian Label Propagation," IEEE Transactions on Image Processing, Vol. 23, No. 3, pp. 1084-1096, March 2014.

[6] X. Liu, D. Zhai, D. Zhao, G. Zhai, and W. Gao, "Progressive image denoising through hybrid graph laplacian regularization: A unified framework," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1491-1503, April 2014.

[7] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE image quality assessment Database Release 2," [Online]. Available: http://live.ece.utexas.edu/research/quality

[8] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008-A database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, pp. 30-45, 2009.

[9] E. C. Larson and D. M. Chandler, "Categorical image quality (CSIQ) database," [Online], Available: http://vision.okstate.edu/csiq

[10] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo, "Color image database TID2013: Peculiarities and preliminary results," *4th European Workshop on Visual Information Processing EUVIP2013*, pp.106-111, June 2013.

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

[12] 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, December 2005.

[13] H. R. Sheikh, and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430-444, February 2006.

[14] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Self-adaptive scale transform for IQA metric," *Proc. IEEE Int. Symp. Circuits and Syst.*, pp. 2365-2368, May 2013.

[15] K. Gu, G. Zhai, X. Yang, W. Zhang, and M. Liu, "Structural similarity weighting for image quality assessment," *Proc. IEEE Int. Conf. Multimedia and Expo Workshops*, pp. 1-6, July 2013.

[16] G. Zhai, X. Wu, X. Yang, W. Lin, and W. Zhang, "A psychovisual quality metric in free-energy principle," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 41-52, January 2012.

[17] K. Gu, G. Zhai, X. Yang, and W. Zhang, "A new reduced-reference image quality assessment using structural degradation model," *Proc. IEEE Int. Symp. Circuits and Syst.*, pp. 1095-1098, May 2013.

[18] K. Gu, G. Zhai, X. Yang, W. Zhang, and M. Liu, "Subjective and objective quality assessment for images with contrast change," *Proc. IEEE Int. Conf. Image Process.*, September 2013.

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

[20] 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, August 2012.

[21] 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, December 2012.

[22] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, pp. 305-312, June 2011.

[23] K. Gu, G. Zhai, X. Yang, W. Zhang, and L. Liang, "No-reference image quality assessment metric by combining free energy theory and structural degradation model," *Proc. IEEE Int. Conf. Multimedia and Expo*, pp. 1-6, July 2013.

[24] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 209-212, March 2013.

[25] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, July 2013.

[26] D. L. Ruderman, "The statistics of natural images," *Network Computation in Neural Syst.*, vol. 5, no. 4, pp. 517-548, 1994.

[27] C-C. Chang and C-J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. Intelligent Systems and Technology*, vol. 2, no. 3, 2011. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[28] Seber, G. A. F. *Multivariate Observations*. Hoboken, NJ: John Wiley & Sons, Inc., 1984.

[29] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527-1554, 2006.

[30] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 1, pp. 52-56, February 1995.

[31] Andrew Ng. UFLDL Tutorial [EB/OL]. http://deeplearning.stanford.edu/wiki/index.php/UFLDL_Tutorial.

[32] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," March 2000, http://www.vqeg.org/.