A Blind Image Quality Metric using a Selection of Relevant Patches based on Convolutional Neural Network

Aladine Chetouani
PRISME Laboratory, University of Orleans
aladine.chetouani@univ-orleans.fr

***Abstract*—Image quality assessment is an important field in different computer vision applications. A plethora of metrics has been proposed in the literature to answer this request. In this paper, we propose an image quality framework without reference based on selection of saliency patches and Convolutional Neural Network. The idea is here to not consider all patches of the distorted image but rather some of them, which are considered as the more perceptually relevant and thus impact more the Mean Opinion Score of the image. To do that, we first compute the saliency map of the distorted image. A scanpath prediction method, that aims to reproduce the visual behavior, is then applied to select the more relevant patches. A Convolutional Neural Network model is finally used to predict the quality score. Its input is the selected patches, while its output is the predicted Mean Opinion Score. The proposed was evaluated using four well-known datasets (LIVE-P2, TID 2008, TID 2013 and CSIQ). The results obtained show its efficiency.**

Keywords—Image quality; CNN model; Saliency; Scanpath prediction

I. Introduction

Due its importance in several computer vision applications, image quality domain became a growing domain [1]. A plethora of methods has been proposed in the literature. We can dissociate three approaches depending on the availability of the reference image. When the latter is accessible, Full Reference (FR) metrics can be used, while No Reference (NR) metrics are used to predict the quality without the reference image. The third approach, often called Reduced Reference (RR), is an alternative solution under the assumption that only some characteristics of the reference image are available.

In this study, we focus on 2D-IQM and we propose a CNN-based blind image quality framework using the saliency information. Neural-based methods have been previously used to predict the quality. In [2], several selected features were selected and combined using a Multi Layer Perceptron (MLP). In [3], a CNN-based blind image quality has been proposed. The image is first decomposed into patches and used as inputs to the CNN model. During the learning step, patches of a given image have the same target and it corresponds to the subjective score (often called MOS: Mean Opinion Score) of the whole image. In [4], a multi-task CNN model is described. The goal was to predict the quality and the degradation type. In [5], a weight computed for each patch was used to estimate the global quality. In [6], the authors discussed about the utilization of deep learning for blind image quality assessment.

The above-cited studies considered all patches of the image and affected to each of them the same MOS. Here, we propose a framework to predict the quality of a given distorted image without reference by considering only the more relevant patches according to the saliency information. The idea developed here is that the global quality of the image is more impacted by saliency regions than the others. So, for a given degraded image, we first compute the corresponding saliency map and relevant patches is then selected by applying a scanpath prediction method. The latter aims to reproduce the behavior of human when the image is analyzed. Selected patches are normalized and used as inputs to a CNN model. The quality of the image is finally given by averaging the predicted scores of the selected patches.

Our paper is organized as follows: In Section 2, the proposed method is described. In Section 3, we present the obtained results in terms of correlation with the subjective judgments. The last section is dedicated to the conclusion and the perspectives.

II. Proposed Method

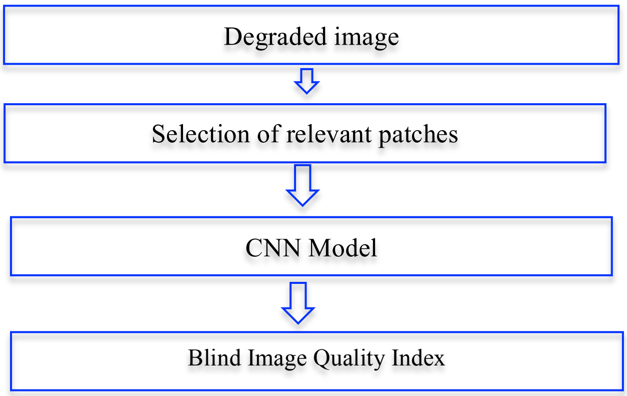


Fig. 1. Flowchart of the proposed method

The flowchart of the proposed method is presented in Fig. 1. For a given distorted image, we first extract specific patches of size 32x32, selected according to the saliency information. These patches are then normalized (mean and standard deviation) and used as inputs to our CNN model. The global estimation quality is then given by the mean of the local-predicted scores. Each of these steps is described in this section.

A. Selection of Relevant Patches

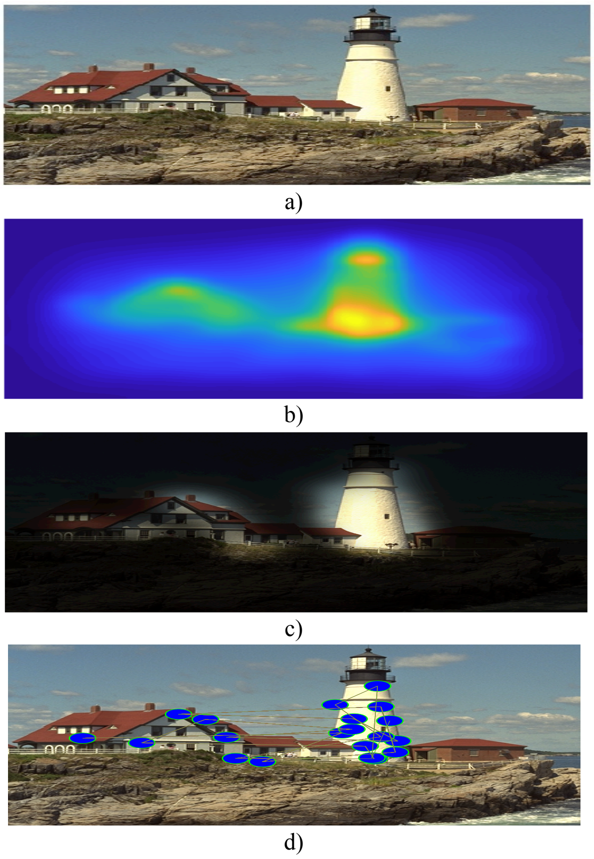
Visual attention plays an important role in image analysis domain and has been exploited in several applications (image retrieval [7], indoor localization [8] and so on). The detected regions, so-called saliency regions, represent the more attractive zones in the image and thus play an important role in the comprehension of the image. Here, we propose to exploit this information by extracting relevant patches and use it as inputs to a CNN model. The underlying idea developed here is that those regions impact highly the subjective judgment and thus the overall quality [9,10]. So, this procedure permits to not consider patches that have a limited impact in terms of quality.

In this study, we used the method proposed in [18] to select the more relevant patches. The latter predicts visual scanpaths of observers based on a saliency model and biases (saccade amplitude and saccade orientation biases). Each position of a given scanpath represents one of the more relevant saliency regions. Fig. 2.d presents a scanpath obtained for the Fig. 2.a. The used method needs two inputs: an image and its corresponding saliency.

The Saliency map can be obtained by environmental characteristics (color, intensity, orientation, etc.) or can be conducted by human observers' deliberate intentions according to some a priori information. The former approach is called "Bottom-up" approach, while the latter is called "Top-down" approach. As the quality evaluation is often driven by the environmental characteristics without prior information (which is the case in this study), a Bottom up method has been used.

One of the first bottom-up models has been developed by Itti et al. [11]. The authors proposed to combine different maps using low-level attributes (intensity, color and orientation). In [12], a more complex method based on some Human Visual System (HVS) characteristics has been described. The saliency map is performed through different perceptual steps (perceptual color space transformation→ Contrast Sensitivity Function filtering [13] → Cortex transform [14] → masking effect → center surround filtering). Other simpler methods have been also proposed. In [15], the authors proposed to determine the saliency of a given image in the Fourier domain. The residual spectrum of the image is first computed and subtracted from its filtered version. The saliency map is then obtained by the inverse Fourier transform. In this work, the Graph-Based Visual Saliency (GBVS) method [16], which is one of the best methods in the state-of-the-art [17], has been used.

In Fig. 2, we show a distorted image (Fig. 2.a), the corresponding saliency map (Fig. 2.b) and the predicted scanpath (Fig. 2.d) where the blue points represent the predicted relevant positions. As we can see, the selected positions are localized on the more attractive regions of the image (lighthouse and houses) and are thus in accordance with our perception.



a)

b)

c)

d)

Fig. 2. Examples of selected patches: a) Distorted image, b) its saliency map c) highlight image and d) the predicted scanpath.

For each predicted position, a patch of size 32x32 is extracted and normalized (i.e. local mean=0 and local standard deviation=1, filter size = 3x3).

B. Architecture of our CNN Model

The selected normalized-patches are then used as inputs to the proposed CNN model, presented in Fig. 3 (32x32x1 → 26x26x16 → 13x13x16 → 7x7x16 → 1x1x48 → 400 → 1). There are two convolutional layers, two pooling layers, one fully connected layer and one output layer. Both convolutional layers are composed of 16 kernels (16 feature maps) of size 7x7. The first pooling layer is a 2x2 min pooling without overlap (i.e. stride=2), while the second pooling layer is composed of three pooling steps (min – max – mean) without overlap. For the latter, the obtained feature maps are pooled to 16 max, 16 min and 16 mean (i.e. each map is pooled to one max, one min and one mean values). This structure permits to better describe the distribution of the kernel maps. The values obtained are then concatenated to form a vector of size 48 (i.e. 16 * 3) and used as inputs to the first fully connected layer of size 400, followed by a dropout step (0.5). The last layer is a logistic regression layer with one output (predicted MOS). As activation function, the well-known ReLu (Rectified Linear Units) has been used after the first fully connected layer.

The parameters fixed during the training step are listed below (the training-validation-test decomposition is presented in section III.A):

- **Start-learning rate:** 0.01
- **Start-momentum:** 0.9
- **End-momentum:** 0.5
- **Optimization method:** Stochastic Gradient Descent (SGD)
- **Test interval (i.e the number of iterations between two evaluation using the validation set):** 1000
- **Batch size:** 64
- **Toolbox:** Caffe [19].

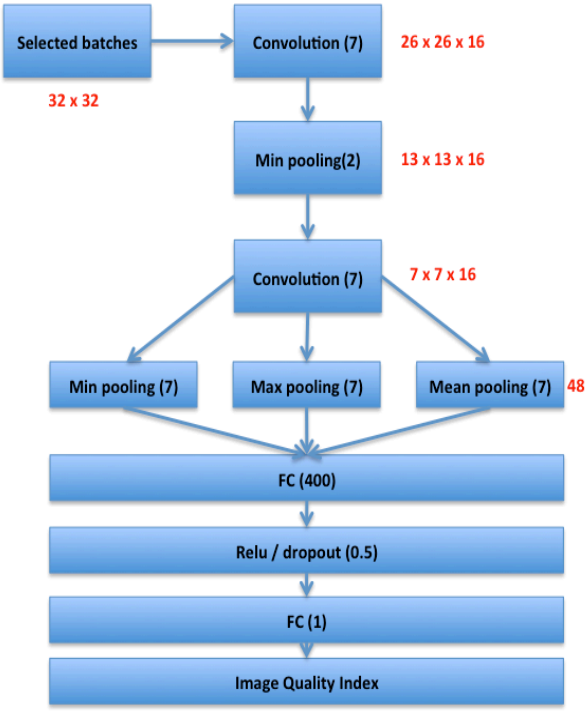


Fig. 3. Architecture of our CNN model

C. *Used Databases*

Four well-known datasets were used to evaluate our method:

- **LIVE image database - Phase 2 (LIVE2-P2) [20]:** this database is composed of 5 degradation types (JPEG2000, JPEG, White Noise, Gaussian Blur and Fast Fading) applied to 29 reference images (779 degraded images). For each degraded image, the DMOS (Differential Mean Opinion Score) is provided.
- **TID 2008 (TID08) [21]:** constituted of 17 degradation types applied to 25 original images, this dataset is composed of 1700 degraded images and its corresponding MOS (Mean Opinion Score).
- **TID 2013 (TID13) [22]:** this dataset is an extended version of the previous one. More degradation types were considered (24 instead of 17) with more degraded images per degradation type (125 instead of 100). So, a total of 3000 degraded images and its corresponding MOS have been provided.
- **CSIQ [23]:** six degradation types are here considered. 866 degraded images are obtained from 30 pristine images. For each of them, the DMOS are given.

III. Experimental Results

A. *Evaluation Protocols*

In order to evaluate the proposed method, two evaluation protocols have been applied:

1. **Protocol 1:** in this protocol, only LIVE-P2 database was used. The latter is decomposed into training-validation (60% for the training and 20% for the validation) and test (20%) sets randomly without overlap. This procedure was repeated 10 times (cross validation). The mean performances are then shown.
2. **Protocol 2:** the objective was here to evaluate the generalization ability of our method. For that, we applied the cross-dataset validation method by using the whole LIVE-P2 dataset to train our CNN model and the others (TID08, TID13 and CSIQ) as test sets.

For both protocols, the Pearson (PCC) and the Spearman (SROCC) coefficient correlations were used to evaluate the capacity of our method to predict subjective judgments. The best performance is represented in bold on grey background.

B. *Evaluation*

1) *Protocol 1: LIVE-P2 dataset*

TABLE I. Protocol 1: Obtained correlations for LIVE-P2 dataset

LIVE-P2			
		PCC	**SROCC**
FR-IQM	**PSNR**	0.856	0.866
	SSIM [24]	0.906	0.913
	FSIM [25]	0.960	0.964
	DeepQA [26]	0.981	0.982
NR-IQM	**DIIVINE [27]**	0.917	0.916
	BLIINDS-II [28]	0.930	0.931
	BRISQUE [29]	0.942	0.940
	CORNIA [30]	0.935	0.942
	IQA-CNN [3]	0.953	0.956
	IQA-CNN+ [4]	0.953	0.953
	IQA-CNN++ [4]	0.950	0.950
	SOM [31]	0.962	0.964
	CNN-Prewitt [32]	0.966	0.958
	Image-wise CNN [6]	0.963	0.964
	Our method	**0.988**	**0.989**

In this section, we present the results obtained for the Protocol 1. The number of patches used during the training is up to 42 000, while the number of patches used during the test

ISBN 978-90-827970-1-5 © EURASIP 2018

is three times less. Table I shows the obtained PCC and SROCC correlations. As we can see, our method outperformed all the compared metrics, especially the CNN-based methods. Note that for the latter methods, the image is decomposed into patches and all of them (with and without a perceptual weighting step) are considered. So, this result highlights the relevance to focus only on the relevant regions of the image.

2) Protocol 2: Cross Dataset Validation

In this section, the results obtained for the protocol 2 are presented. The number of patches used during the training-validation step is up to 53 000 (i.e. all patches of the LIVE-P2 dataset). Tables II-IV show respectively the performances obtained for TID08, TID13 and CSIQ datasets.

For the TID08 dataset, our method outperformed all the NR metrics and some of the FR measures. The best PCC value was obtained by the FSIM metric, while our method achieved the best SROCC value. Comparing to the CNN-based methods, the performances are close but less patches were used in our case.

TABLE II. PROTOCOL 2: OBTAINED CORRELATIONS FOR TID08 DATASET

TID 2008			
		PCC	**SROCC**
FR-IQM	**PSNR**	0.776	0.901
	SSIM	0.817	0.903
	FSIM	**0.952**	0.954
NR-IQM	**CORNIA**	0.890	0.880
	IQA-CNN	0.903	0.920
	IQA-CNN+	0.893	0.912
	IQA-CNN++	0.895	0.906
	SOM	0.899	0.923
	Our method	0.91	**0.956**

For the TID13 dataset, our method outperformed also all the compared NR-IQM and is competitive with the DeepQA, which is a FR metric.

TABLE III. PROTOCOL 2: OBTAINED CORRELATIONS FOR TID13 DATASET

TID 2013			
		PCC	**SROCC**
FR-IQM	**PSNR**	0.706	0.636
	SSIM	0.691	0.775
	DeepQA	**0.946**	0.940
NR-IQM	**CORNIA**	0.613	0.549
	BRISQUE	0.651	0.572
	Image-wise CNN	0.802	0.800
	Our method	0.925	**0.955**

For the CSIQ dataset, we obtained the best results whatever the kind of metrics (NR and FR). Comparing to the CNN-based methods, the achieved PCC and SROCC values are higher than those metrics.

TABLE IV. PROTOCOL 2: OBTAINED CORRELATIONS FOR CSIQ DATASET

CSIQ			
		PCC	**SROCC**
FR-IQM	**PSNR**	0.800	0.806
	SSIM	0.861	0.876
	FSIM	0.961	0.962
	DeepQA	0.964	0.960
NR-IQM	**BRISQUE**	0.797	0.756
	CORNIA	0.914	0.899
	IQA-CNN	0.903	0.923
	IQA-CNN+	0.910	0.918
	IQA-CNN++	0.928	0.936
	Our method	**0.968**	**0.973**

In Fig. 4, we show the predicted MOS vs. MOS for the test datasets (TID08, TID13 and CSIQ). The scatter plot of the data represents visually the correlation between the subjective scores and its predicted version (objective scores). The red curve corresponds to the logistic function obtained by interpolating the objective scores. As we can see, the scatter distributions are consistent, especially for the CSIQ dataset for which the best performances were obtained.

IV. CONCLUSION

In this paper, we proposed a method to estimate the quality of 2D distorted images by selecting some relevant patches as inputs to a CNN model. The selection of these patches was realized using a visual scanpath prediction method that exploits the saliency information. The results obtained are compared to the state-of-the-art and show the relevance of our approach. As perspective, we will try to use different saliency methods and compare the performance variations according to this input.

REFERENCES

[1] P. Grother, and E. Tabassi, “Performance of Biometric Quality Measures”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.29, no.4, pp. 531-543, 2007.

[2] A. Chetouani, A. Beghdadi, S. Chen, and G. Mostafaoui, “A free reference image quality measure using neural networks”, International Workshop on Video Processing and Quality Metrics, 2010.

[3] L. Kang, P. Ye, Y. Li and D. Doermann, "Convolutional Neural Networks for No-Reference Image Quality Assessment," 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1733-1740, 2014

[4] L. Kang, P. Ye, Y. Li and D. Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," 2015 IEEE International Conference on Image Processing, pp. 2791-2795, 2015

[5] S. Bianco, L. Celona, P. Napoletano and R. Schettini, “On the Use of Deep Learning for Blind Image Quality Assessment”, Computer Vision and Pattern Recognition, 2016.

[6] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang and A. C. Bovik, "Deep Convolutional Neural Models for Picture-Quality Prediction: Challenges and Solutions to Data-Driven Image Quality Assessment," in IEEE Signal Processing Magazine*,* vol. 34, no. 6, pp. 130-141, Nov. 2017.

[7] C. A. Hussain, D. V. Rao and S. A. Masthani, “Robust Pre-processing Technique Based on Saliency Detection for Content Based Image Retrieval Systems”, In Procedia Computer Science, Volume 85, pp. 571-580, 2016.

[8] W. Elloumi, K. Guissous, A. Chetouani and S. Treuillet, “Improving a vision indoor localization system by a saliency-guided detection”, IEEE VCIP pp. 149-152, 2014.

[9] D. V. Rao, N. Sudhakar, I. R. Babu, and L. P. Reddy, “Image quality assessment complemented with visual region of interest,” in Proc. Int. Conf. Comput.: Theory Applicat., pp. 681–687, 2007.

[10] Q. Ma and L. Zhang, “Image quality assessment with visual attention,” in Proc. ICPR, pp. 1–4, 2008.

[11] Itti L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11): p. 1254-1259, 1998.

[12] Lemeur O. Le Meur, P. Le Callet, D. Barba and D. Thoreau, “A coherent computational approach to model the bottom-up visual attention”, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 28, N°5, 2006.

[13] CSF A.B. Watson, “Visual detection of spatial contrast patterns: Evaluation of five simple models”, Optics Express, pp.12–33, 2000.

[14] Cortex A.B. Watson. The Cortex transform: rapid computation of simulated neural images. Computer Vision Graphics and Image Processing*,* pp. 311–327, 1987.

[15] X. Hou, L. Zhang, “Saliency Detection: A Spectral Residual Approach”, Computer Vision and Pattern Recognition, 2007.

[16] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency", Proceedings of Neural Information Processing Systems (NIPS), 2006.

[17] http://saliency.mit.edu/

[18] 0. Le Meur and Liu Z., “Saccadic model of eye movements for free-viewing condition”, *Vision Research*, 2015.

[19] http://caffe.berkeleyvision.org/

[20] H.R. Sheikh, Z.Wang, L. Cormack and A.C. Bovik, "LIVE Image Quality Assessment Database Release 2", http://live.ece.utexas.edu/research/quality.

[21] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, F. Battisti, "TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics", Advances of Modern Radioelectronics, Vol. 10, pp. 30-45, 2009.

[22] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.-C. Jay Kuo, Image database TID2013: Peculiarities, results and perspectives, Signal Processing: Image Communication, vol. 30, pp. 57-77, 2015.

[23] E. C. Larson and D. M. Chandler, "Most Apparent Distortion: Full-Reference Image Quality Assessment and the Role of Strategy," Journal of Electronic Imaging, 19 (1), March 2010.

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

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

[26] J. Kim and S. Lee, "Deep Learning of Human Visual Sensitivity in Image Quality Assessment Framework," IEEE Conference on Computer Vision and Pattern Recognition, pp. 1969-1977, 2017.

[27] A. K. Moorthy and A. C. Bovik, "Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality", IEEE Transactions on Image Processing, 2011.

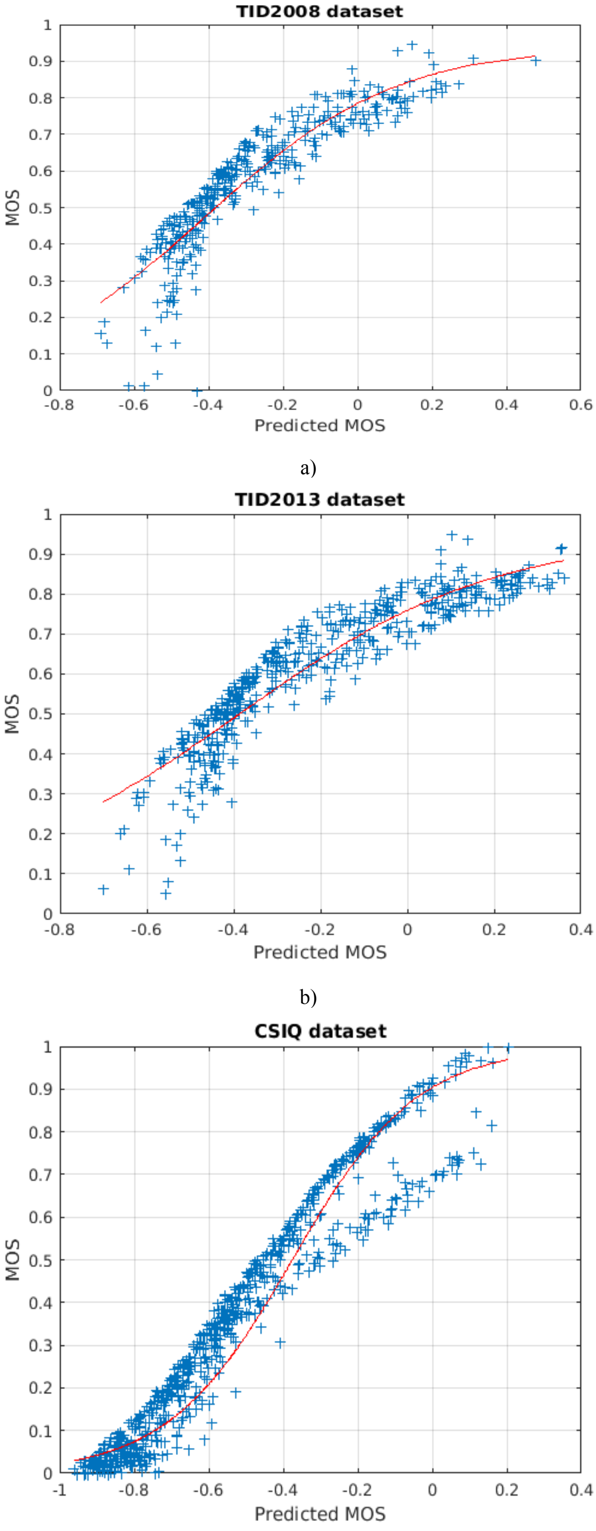
[28] M.A Saad and A. C. Bovik, "Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain, " IEEE Transactions on Image Processing, pp. 1, 2012.

[29] A. Mittal, A. K. Moorthy and A. C. Bovik, “Referenceless Image Spatial Quality Evaluation Engine,” 45th Asilomar Conference on Signals, Systems and Computers, November 2011

[30] P. Ye, J. Kumar, L. Kang and D. Doermann, "Unsupervised Feature Learning Framework for No-reference Image Quality Assessment", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1098-1105, 2012

[31] Peng Zhang, Wengang Zhou, Lei Wu and Houqiang Li, "SOM: Semantic obviousness metric for image quality assessment," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2394-2402, 2015.

[32] J. Li, L. Zou, J. Yan, D. Deng, T. Qu, and G. Xie, “No-reference image quality assessment using Prewitt magnitude based on convolutional neural networks”, Signal, Image and Video Processing. Vol. 10, 10.1007/s11760-015-0784-2, 2015.



a)

b)

c)

Fig. 4. Predicted MOS vs MOS: a) TID08, TID13 and c) CSIQ datasets