Grouping strategies to improve the correlation between subjective and objective image quality data

Silvia Corchs, Francesca Gasparini, and Raimondo Schettini

Department of Informatics, Systems and Communication, University of Milano-Bicocca,
Viale Sarca 336, 20126 Milano, Italy

ABSTRACT

In this paper we address the problem of Image Quality Assessment of no reference metrics, focusing on JPEG corrupted images. In general no reference metrics are not able to measure with the same performance the distortions within their possible range and with respect to different image contents. The crosstalk between content and distortion signals influences the human perception. We here propose two strategies to improve the correlation between subjective and objective quality data. The first strategy is based on grouping the images according to their spatial complexity. The second one is based on a frequency analysis. Both the strategies are tested on two databases available in the literature. The results show an improvement in the correlations between no reference metrics and psycho-visual data, evaluated in terms of the Pearson Correlation Coefficient.

Keywords: image quality assessment, no reference metrics, JPEG

1. INTRODUCTION

Image Quality Assessment (IQA) is a very active topic of research.[1] In general, IQA methods can be categorized as subjective versus objective ones.[2,3] Subjective methods are based on psychological experiments involving human observers. Objective image quality approaches can be categorized into three groups depending on the availability of the original image. Full Reference (FR) methods perform a direct comparison between the image under test and a reference or original image. No Reference (NR) metrics, also called blind methods, are applied when the original image is unavailable. Reduced Reference (RR) metrics lie between FR and NR metrics and are designed to predict image quality with only partial information about the reference image.[1]

It is in general assumed that subjective methods produce an actual estimate of the perceived quality while objective methods produce values that should be correlated with human perceptions as best as possible. Great effort is devoted by the image quality community to proper correlate objective and subjective data. In general logistic and polynomial regressions are used. Objective and subjective results can be compared through different performance measures. Typical measures of performance are the Pearson Correlation Coefficient (PCC), Spearman Rank Order Correlation Coefficient (SROCC) and Kendall Rank Order correlation.

In this paper we aim to improve the agreement between objective and subjective data focusing on two different strategies. The first one is based on grouping the images according to their spatial complexity. The second strategy we present here is based on the analysis of the spectral frequency of the images. In this article we focus on JPEG corrupted images. Six NR Blockiness metrics and three general purpose blind metrics are considered.

Different benchmark databases are available to test the algorithms' performance with respect to the human subjective judgments. Among the most frequently used, we have here chosen LIVE[4] and CSIQ[5] databases. The paper is organized as follows: in section 2 the NR metrics here considered. The two proposed strategies are described in sections 3 and 4 respectively. In section 5 the results of applying both strategies to images of LIVE[4] and CSIQ[5] databases are presented and the performances in terms of Linear Pearson Correlation coefficients are reported. Finally, the conclusions are drawn in 6.

Send correspondence to: Francesca Gasparini E-mail: francesca.gasparini@disco.unimib.it, Telephone: 0039 0264487856

Image Quality and System Performance X, edited by Peter D. Burns, Sophie Triantaphillidou,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 8653, 86530D • © 2013 SPIE-IS&T
CCC code: 0277-786X/13/$18 • doi: 10.1117/12.2006227

2. NO REFERENCE METRICS

Several No Reference metrics exist in the literature. In this work we have considered six NR blockiness metrics (1-6) and three NR general purpose metrics (7-9):

1. **PAN**, developed by Pan et al.:[6] It is based on gradient features. It examines the blocks individually, measuring the severity of blocking artifacts locally. The local metric is averaged over all possible blocks to yield a unique score. It takes into account the blocking artifacts for high bit rate images and the flatness for the very low bit rate images.

2. **VLA**, developed by Vlachos:[7] The method operates in the frequency domain using fast transformations. It uses the cross-correlation of subsample images to detect and measure blocking artifacts. The summation of the phase correlations between some sets of sub-images, which measures the intra-block similarity, is divided by the summation of the phase correlations between some other sets of sub-images, which measures the inter-block similarity, to yield a measure of image blockiness.

3. **WSB**, developed by Wang et al.:[8] The method considers blurring and blocking as the most significant artifacts generated during the JPEG compression process. It works in the frequency domain and is based on gradient features, that are combined to constitute a quality prediction model.

4. **WBE**, developed by Wang et al.:[9] The key idea is to model the blocky image as a non-blocky image interfered with a pure blocky signal. The method is formulated in the frequency domain. The task of the blocking effect measurement algorithm is then to detect and evaluate the peaks in the power spectrum of the blocky signal. Luminance and texture masking effects are taken into account within the metric.

5. **GBIM**, Generalized Block-edge Impairment Metric developed by Wu and Yuen:[10] It is the most well known metric in the spatial domain.The GBIM assumes that the artifacts occur on a grid of blocks of pixels, which is common for most compression standards. It measures the blockiness separately in horizontal and vertical directions,and then combined them into a single quality value.

6. **CHEN** developed by Chen and Bloom:[11] For a given image, the absolute difference between horizontally adjacent pixels is computed, normalized, and averaged along each column. A one-dimensional discrete Fourier transform is thereafter applied and a vertical blockiness measure is derived. The horizontal measure is computed similarly. Finally, the blockiness measure for the given image is formulated by pooling the two directional measures.

7. **BIQI** (Blind Image Quality Indices) developed by Moorthy and Bovik:[12] It is a two-step framework for general purpose No Reference Image Quality Assessment, based on natural scene statistics (NSS). Once trained, the framework does not require any knowledge of the distorting process and the framework is modular in that it can be extended to any number of distortions.

8. **BRISQUE** (Blind/Referenceless Image Spatial QUality Evaluator) developed by Mittal et al.:[13] It is a NR general purpose metric based on natural scene statistic (NSS) which operates in the spatial domain. It uses scene statistics of locally normalized luminance coefficients to quantify possible losses of naturalness in the image due to the presence of distortions, leading to a holistic measure of quality.

9. **BLIINDS** (BLind Image Integrity Notator using DCT-Statistics) developed by Saad et al.[14] It is a general-purpose metric that uses natural scene statistics models of discrete cosine transform (DCT) coefficients to perform distortion-agnostic NR Image Quality Assessment.

In general NR-IQ metrics are not able to measure with the same performance the actual distortions disregarding the image contents. Considering different blockiness metrics separately and focusing on a single image (i.e. single content), a monotone behavior is observed as the blockiness increases. In Figure 1 top, the absolute value of the WBE metric is plotted as function of the bit per pixel ratio for five example images. If different image contents are considered, a non monotone profile for each metric is often found. This behavior is due to

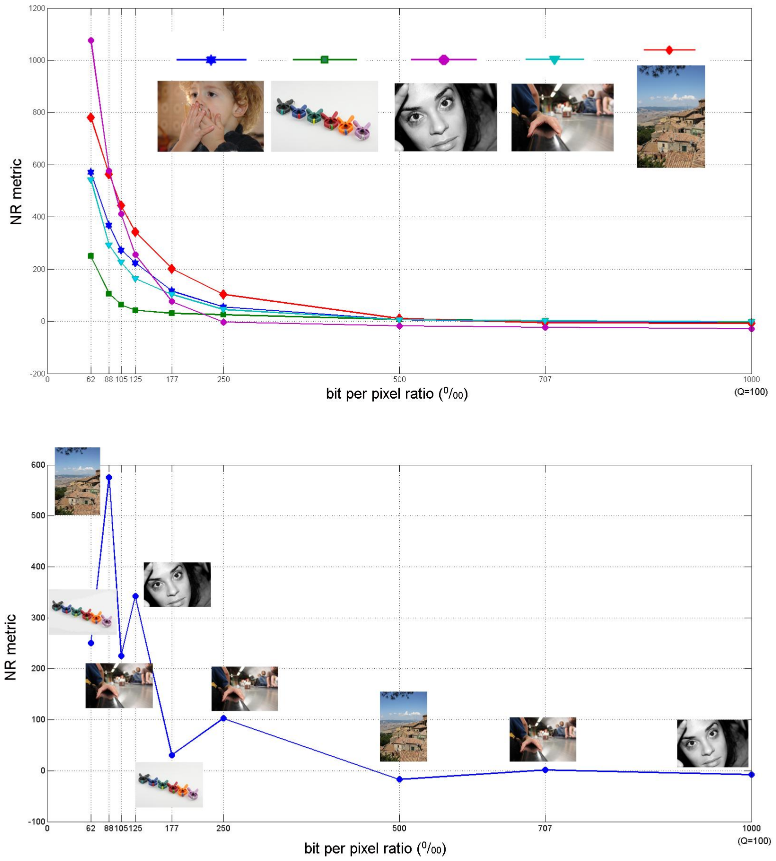

Figure 1. Top: the WBE NR metric[9] applied to 5 different images with 9 levels of JPEG compression. Bottom: the same metric applied to different contents.

the fact that different signal contents influence differently the measure of the same level of distortion, (Figure 1 bottom).

Keeping these facts in mind, we here propose two strategies to improve the correlation of subjective and objective quality data that are hereafter described.

3. STRATEGY BASED ON THE IMAGE COMPLEXITY INDEX

In order to evaluate the image complexity, we have considered the index proposed by Chacon and Corral.[15] These authors proposed a fuzzy approach to determine the complexity of an image, based on the analysis of the edge level percentages. Three classes of images are determined: low complexity images, medium complexity images, and high complexity ones. In Figure 2 we show the classification of the 29 original images of the LIVE database in terms of complexity.

In Figure 3 the values of the WSB metric applied to all the images of the LIVE databases are shown. Each curve corresponds to a single image content. In the figure on the left, all the 29 images of the LIVE database with their corresponding distorted versions are reported. Blue curves correspond to images belonging to the low complexity group, red curves to images of the medium complexity one, while green ones to high complexity images. The other three plots report separately the WBE metrics evaluated for images of each subgroups from low complexity to high complexity respectively.

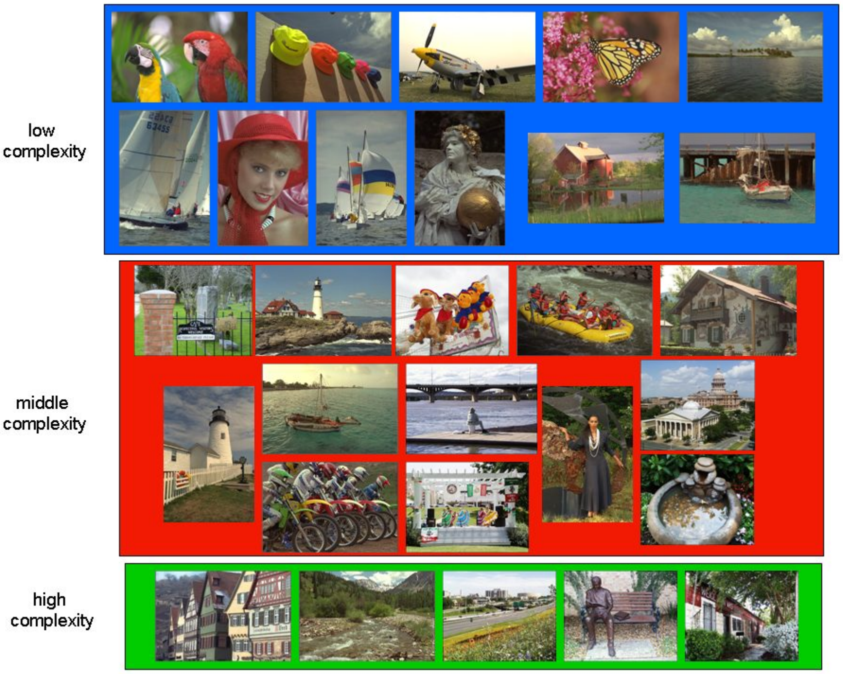

Figure 2. Images of the LIVE database, grouped with respect to complexity (top images: low complexity; middle images: medium complexity; bottom images: high complexity.

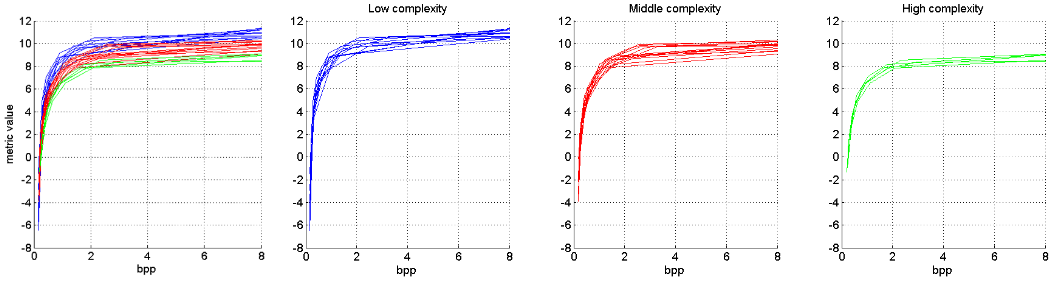

Figure 3. Left image: WSB metric applied to all the JPEG distorted images of the LIVE databases: blue curves correspond to images belonging to the low complexity group, red curves to images of the medium complexity one, while green ones to high complexity images. The other three plots report separately the WBE metrics evaluated for images of each subgroups from low complexity to high complexity respectively. Note that each curve corresponds to a single image content.

4. STRATEGY BASED ON THE SPECTRAL FREQUENCY ANALYSIS

We compute the frequency in the Fourier domain, corresponding to the 99% of the image energy. Each one of the nine metrics considered are weighted by a frequency-dependent factor, obtained normalizing this frequency with respect to the Nyquist frequency. In Figure 4 the original images of the LIVE database are sorted with respect to increasing frequency, starting from the top left corner, to the bottom right one.

5. RESULTS

In this section we present and compare the experimental results obtained for the two databases: LIVE (233 images)[4] and CSIQ (150 images).[5]

If we consider a single metric and the corresponding subjective scores for a given database, a logistic regression curve can be computed. We will refer to this regression as R_a, that is the function obtained using all the data as it is usually done in the literature.

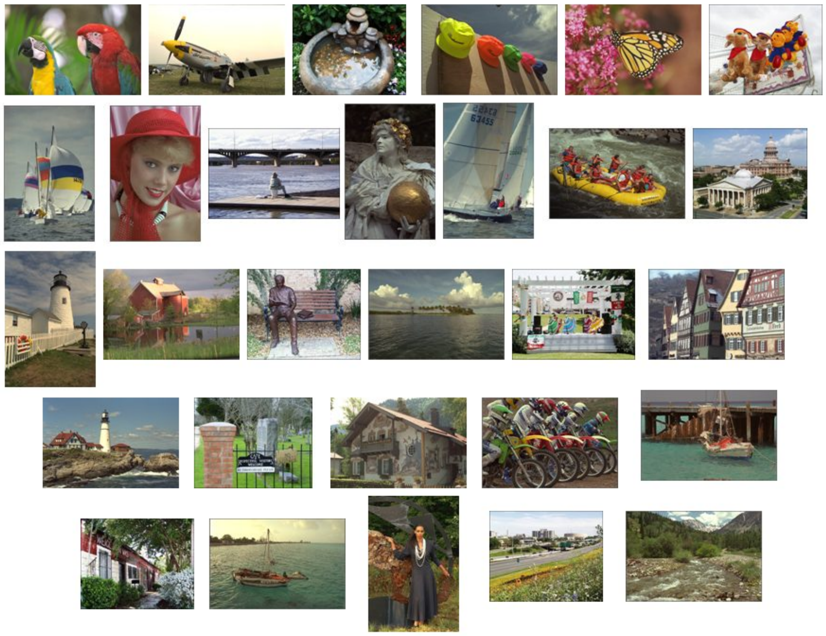

Figure 4. The 29 reference images of the LIVE database sorted with respect to increasing frequency, starting from the top left corner, to the bottom right one.

With respect to strategy 1 (complexity-based), we have first evaluated the complexity index for each of the distorted images. We have thus obtained the high, medium and low complexity-groups. Within each of the groups, the regression between each of the objective metrics and the corresponding subjective scores is performed. Hereafter we refer to these regression functions as R_H, R_M and R_L respectively. For instance, in figure 5 the regressions of the WBE metric with respect to the subjective scores on the CSIQ database are considered. In particular: top row, first column corresponds to the R_a regression; top row, second column, corresponds to the R_L regression; bottom row, first column, corresponds to the R_M regression; while bottom row, second column, corresponds to the R_H regression.

With respect to strategy 2 (frequency-based), we have first weighted each metric with the frequency-based factor. We have then correlated this modified objective metric and the corresponding subjective scores for all the data of a given database. The regression between the subjective scores obtained by the psycho-visual experiment and the quality predicted by the frequency-weighted metric is here called R_f.

In order to compare the performance of the different regression curves (corresponding to the different strategies here proposed), we have evaluated their PCC, that is the linear correlation coefficient between the quality predicted by a metric and the subjective scores. The PCCs obtained with respect to the different strategies here proposed are reported in Tables 1 and 2, where the notation used is as follows:

- PCC(R_a, a): PCC obtained using the R_a regression function and all the images of a given dataset;
- PCC(R_H, H): PCC obtained using R_H regression function and the images corresponding to the high-complexity group;
- PCC(R_M, M): PCC obtained using R_M regression function and the images corresponding to the medium-complexity group;
- PCC(R_L, L): PCC obtained using R_L regression function and the images corresponding to the low-complexity group;

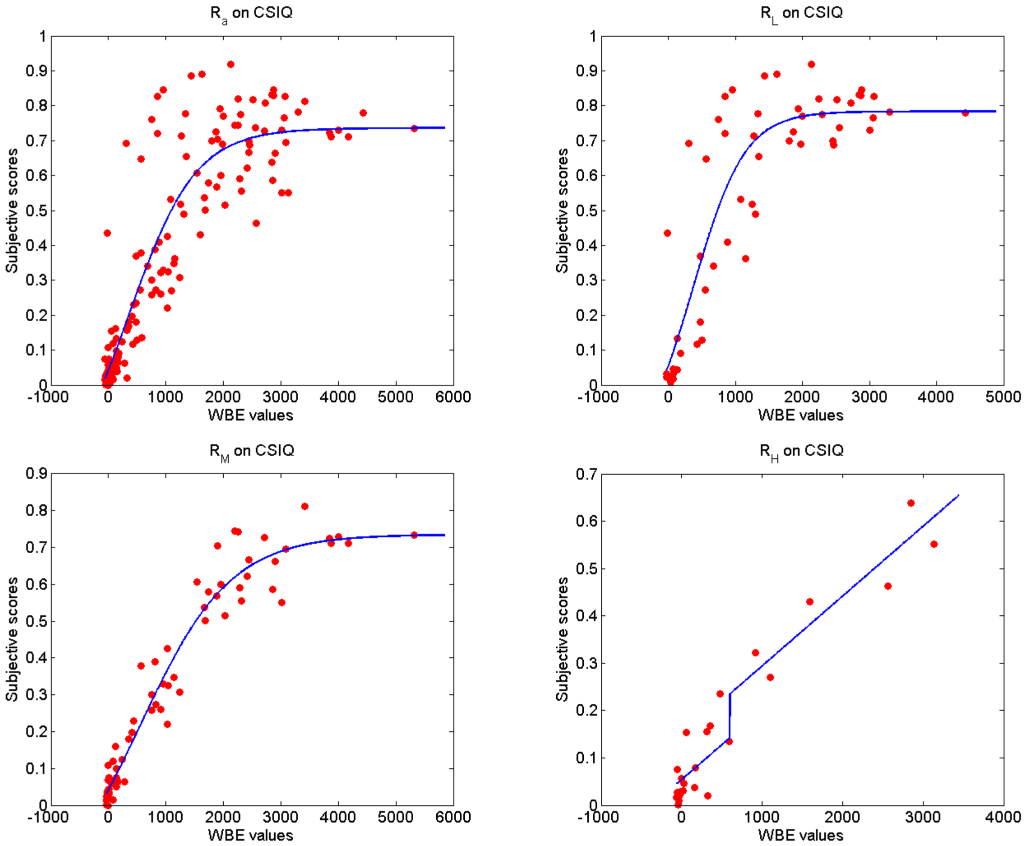

Figure 5. The regressions of the WBE metric with respect to the subjective scores on the CSIQ database are reported. Top row, first column R_a; top row, second column: R_L ; bottom row, first column: R_M ; bottom row, second column: R_H.

- PCC(R_a, H): PCC obtained using R_a regression function and the images corresponding to the high-complexity group;
- PCC(R_a, M): PCC obtained using R_a regression function and the images corresponding to the medium-complexity group;
- PCC(R_a, L): PCC obtained using R_a regression function and the images corresponding to the low-complexity group;
- PCC(R_f, a): PCC obtained using the R_f regression function and all the images of a given dataset.

In the tables, the bold characters indicate the best performance achieved for each of the single metrics. For the LIVE database, the grouping strategy improve the results for all the metrics for the case of medium and high complexities while for the low complexity group the improvements are obtained for six of the nine metrics considered. For the CSIQ database, the grouping strategy improves the performances in all the metrics and for all the three complexity groups except for only one metric (in the high complexity case). With respect to the frequency-based strategy, improvements have also been obtained for some of the metrics.

6. CONCLUSIONS

In this work we have proposed two different strategies in order to improve the correlation between subjective and objective image quality data. We have focussed on JPEG-corrupted images. Several no reference metrics for blockiness and general purpose ones have been considered. The first strategy is based on the analysis of the image

Metric		*Freq-Weight*	*Grouped by complexity*					
	PCC(R_a,a)	PCC(R_f,a)	PCC(R_a,L)	PCC(R_L,L)	PCC(R_a,M)	PCC(R_M,M)	PCC(R_a,H)	PCC(R_H,H)
PAN	0.8599	**0.8628**	0.9033	**0.9113**	0.8735	**0.8742**	0.8189	**0.8293**
VLA	0.8307	**0.8534**	0.8708	0.8182	0.7924	**0.7964**	0.8103	**0.8394**
WSB	0.9414	**0.9438**	0.9586	**0.9689**	0.9442	**0.9496**	0.9530	**0.9584**
WBE	0.8304	**0.9356**	0.8325	**0.9546**	0.8366	**0.9268**	0.8356	**0.8565**
GBIM	0.9485	**0.9523**	0.9568	**0.9575**	0.9425	**0.9426**	0.9476	**0.9524**
CHEN	0.9447	**0.9459**	0.9545	**0.9588**	0.9458	**0.9461**	0.9474	**0.9498**
BIQI	0.9181	0.9064	0.9211	**0.9397**	0.9257	**0.9275**	0.9400	**0.9428**
BRISQUE	0.9345	0.9329	0.9517	0.9130	0.9261	**0.9268**	0.9053	**0.9124**
BLIINDE	0.9105	**0.9127**	0.9308	0.9293	0.9157	**0.9337**	0.8933	**0.9393**

Table 1. The PCCs with respect to the strategies proposed for the LIVE database.

Metric		*Freq-Weight*	*Grouped by complexity*					
	PCC(R_a,a)	PCC(R_f,a)	PCC(R_a,L)	PCC(R_L,L)	PCC(R_a,M)	PCC(R_M,M)	PCC(R_a,H)	PCC(R_H,H)
PAN	0.8739	**0.8917**	0.8123	**0.8543**	0.9394	**0.9409**	0.8033	**0.9046**
VLA	0.9063	0.8693	0.8602	**0.8822**	0.9221	**0.9323**	0.8239	**0.8838**
WSB	0.9483	0.9412	0.8941	**0.9000**	0.9791	**0.9808**	0.9748	**0.9749**
WBE	0.9105	0.9091	0.8819	**0.8899**	0.9687	**0.9767**	0.9559	**0.9606**
WSB	0.9348	**0.9429**	0.8747	**0.8881**	0.9684	**0.9710**	0.9610	**0.9681**
GBIM	0.9186	**0.9267**	0.8556	**0.8792**	0.9715	**0.9723**	0.9263	**0.9395**
BIQI	0.8475	0.8438	0.8465	**0.8470**	0.8316	**0.8400**	0.7919	0.7716
BRISQUE	0.9086	**0.9149**	0.8277	**0.8435**	0.9501	**0.9524**	0.9032	**0.9176**
BLIINDE	0.8926	**0.9193**	0.7762	**0.8569**	0.9537	**0.9623**	0.9084	**0.9298**

Table 2. The PCCs with respect to the strategies proposed for the CSIQ database.

complexity while the second one is based on the frequency analysis of the images. Both strategies have been tested on two databases available in the literature. The experimental results confirm that improvements can be achieved if the images are first grouped according to their complexity and then correlated to the corresponding subjective scores (strategy 1). Also applying strategy 2 better results in terms of correlation are obtained.

7. ACKNOWLEDGEMENTS

This project was founded by Océ-Canon.[16]

REFERENCES

[1] A. Bovik and Z. Wang, *Modern Image Quality Assessment.* Claypool Publishers, 2006.

[2] ITU, "Final report from the video quality experts group on the validation of objective models of video quality," tech. rep., ITU-T Study Group 9 Contribution 80, 2000.

[3] ITU, "Methodology for the subjective assessment of the quality for television pictures," tech. rep., ITU-R Rec. BT. 500-11, 2002.

[4] H. Sheik, Z. Wang, L. Cormakc, and A. Bovik, *LIVE Image Quality Assessment Database Release 2.* http://live.ece.utexas.edu/research/quality.

[5] 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), p. 011006, 2010.

[6] F. Pan, S. Lin, S. Rahardja, W. Lin, E. Ong, S. Yao, Z. Lu, and X. Yang, "Locally-adaptive algorithm for measuring blocking artifacts in images and videos," in *Proceedings of the International Symposium on Circuits and Systems*, **3**, pp. 925–928, IEEE, 2004.

[7] T. Vlachos, "Detection of blocking artifacts in compressed video," *IET Electronics Letters* **36**, pp. 1106–1108, 2000.

[8] Z. Wang, H. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of jpeg compressed images," in *Proc. International Conference on Image Processing*, **1**, pp. 477–480, IEEE, 2002.

[9] Z. Wang, A. C. Bovik, and B. L. Evans, "Blind measurement of blocking artifacts in images," in *Proc. International Conference on Image Processing*, **3**, pp. 981–984, IEEE, 2000.

[10] H. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Processing Letters* **4**, pp. 317–320, 1997.

[11] C. Chen and A. Bloom, "A blind reference-free blockiness measure," in *Lecture Notes in Computer Science*, **6297**, pp. 112–123, Springer-Verlag Berlin Heidelberg, 2010.

[12] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters* **17**, pp. 513–516, 2010.

[13] A. Mittal, A. Moorthy, and A. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing* **21**(2), pp. 4695–4708, 2012.

[14] M. Saad, A. Bovik, and C. Charrier, "Model-based blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE Transactions on Image Processing* **21**(8), pp. 3339–3352, 2012.

[15] M. Chacon and A. Corral, "Image complexity measure: a human criterion free approach," in *Proc. Meeting of the North American Fuzzy Information Processing Society*, pp. 241–246, IEEE, 2005.

[16] "http://www.oce.com/," 2012.