LEARNING SPARSE MODELS FOR IMAGE QUALITY ASSESSMENT

Tanaya Guha, Ehsan Nezhadarya, Rabab K Ward

Electrical and Computer Engineering
The University of British Columbia, Vancouver, BC, Canada

ABSTRACT

Many successful image quality metrics rely on the structural information in an image to assess its perceptual quality. Extracting the structural information that is perceptually meaningful to our visual system, however, is a challenging task. This paper proposes a new quality assessment metric that relies on a sparse modeling approach to learn the inherent structures of the image. These structures are learnt as a set of basis vectors, such that any structure in the image can be represented by a linear combination of only a few of these basis vectors. This strategy is known to generate basis vectors that are qualitatively similar to the receptive field of the simple cells present in the mammalian primary visual cortex. The perceptual quality of the distorted image is estimated by comparing the structures of the reference and the distorted images in terms of the learnt basis vectors. Our approach is evaluated on five standard subject-rated image quality assessment datasets. The proposed metric exhibits high correlation with the subjective ratings outperforming several well established methods.

Index Terms— image quality assessment, overcomplete dictionary, sparse representation.

1. INTRODUCTION

The visual quality of an image is often degraded during its acquisition, compression, transmission, storage or reconstruction. Automatic estimation of the quality of an image is important to many image and multimedia processing systems involving monitoring, controlling and improving the quality of the image. The *mean squared error* (MSE) and the *peak signal-to-noise ratio* (PSNR) have been traditionally used to measure the image quality degradations. These metrics are mathematically convenient to use but they do not correlate well with human perception of image quality.

In recent years, a considerable amount of research effort has been put towards quantifying the quality of images as *perceived* by humans, and a number of *objective* image quality assessment algorithms that agree with the subjective judgment of human beings have been developed. The objective quality assessment methods are broadly classified into three categories depending on how much information about the original undistorted image they use. These categories are: no-reference, reduced-reference and full-reference. This paper concentrates on the *full-reference* quality estimation approach.

Modern image quality estimation methods attempt to model the visual content of images based on certain significant properties of the *human visual system* (HVS). One popular approach operates on the basis of an important aspect of the *human visual system* (HVS) - its sensitivity towards the image structures for developing cognitive understanding. In this approach, the perceptual quality of a given distorted image is estimated by comparing its structures with those in its reference image.

An image quality metric which is representative of the class of structural information-based metrics is the *structural similarity index* (SSIM) [1]. This method measures the visual quality of a patch in the distorted image by comparing it with the corresponding patch in the reference image, in terms of three components: luminance, contrast and structure. A global quality score is computed by combining the effects of the three components over all the image patches. SSIM achieved much success because of its simplicity, and its ability to tackle a wide variety of distortions. Due to its pixel-domain implementation, SSIM is highly sensitive to geometric distortions like scaling, translation, rotation and other misalignments. To improve the performance of SSIM, multiscale extension [2], wavelet transform-based modification [3], gradient-domain implementation [4] and various pooling strategies [5] have been proposed.

The key to the success of any structural information-based method lies in its efficiency in capturing the structures from the images. But obtaining or analyzing the structural information in a perceptually meaningful way is a non-trivial task. A widely used mathematical tool for analyzing image structures is the wavelet transform. The wavelet transform however uses a set of predefined, data-independent basis functions, and its success is therefore limited by the degree to how suitable these functions are in capturing the structures of the signals under consideration.

In this paper, we use a more generalized approach, based on learning the sparse representation of the image, to analyze the image structures, and develop a full-reference image quality assessment metric. The proposed method involves learn-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

ing an overcomplete dictionary - a set of basis vectors (that are not necessarily orthogonal), where the number of bases is greater than the dimensionality of the input. Such a dictionary is learnt from the reference image so as to extract the images inherent underlying structural information. With the help of this dictionary, the local structures in the distorted image is analyzed and the perceptual quality of the distorted image is quantified using the resulting sparse coefficients . Our method is evaluated on five publicly available subject-rated image quality assessment databases. Highly encouraging results were achieved on all databases.

2. THE PROPOSED IMAGE QUALITY ASSESSMENT METHOD

The proposed image quality assessment method consists of two steps: the *dictionary learning* step and the *quality assessment* step. In the dictionary learning step, the inherent structures are extracted from the reference image by learning a dictionary from the image itself. In the quality assessment step, the structures in the distorted image are compared with those in the reference image using the previously learnt dictionary. These two steps are described below in detail.

2.1. Dictionary Learning

Given a reference image $I_r \in \mathbb{R}^N$, our first task is to learn an overcomplete dictionary $\mathbf{\Phi}$. This can be achieved by fitting the basis vectors of the dictionary to a collection of (distinct, possibly overlapping) patches extracted from the image itself. These image patches account for the local structures in that image. We first extract a large number of *random* patches, each of dimension $\sqrt{n} \times \sqrt{n}$, then the patches with low or no structural information (homogeneous patches) are discarded. This is accomplished by removing the patches whose variance is zero or close to zero after mean removal. Each of the remaining image patch is then converted to a vector of length n. These patch vectors are concatenated to form a matrix $\mathbf{P} \in \mathbb{R}^{n \times k}$ where k is the number of patches extracted from I_r.

Using these patches in $\mathbf{P}$ as input, we now train an overcomplete dictionary $\mathbf{\Phi} \in \mathbb{R}^{n \times m}$ having m basis vectors where $n < m$. The goal is to represent any patch vector in $\mathbf{P}$ as a linear superposition of no more than τ dictionary columns where $\tau << m$. This is formally written as the following sparse optimization problem:

$$\min_{\{\mathbf{\Phi},\mathbf{X}\}} \quad \left\{ \|\mathbf{P} - \mathbf{\Phi}\mathbf{X}\|_F^2 \right\} \quad \text{s. t. } \forall i \quad \|\mathbf{x}_i\|_0 \leq \tau_1 \qquad (1)$$

where $\|.\|_F$ is the Frobenius matrix norm (square root of the sum of the squared values of all elements in a matrix) and $\|.\|_0$ is the ℓ_0 semi-norm (counts the number of non-zero elements in a vector). The ℓ_0 semi-norm though provides a straightforward notion of sparsity, but it renders the problem non-convex. Thus obtaining an accurate solution of (1) is NP hard. Nevertheless, in the last decade researchers have found practical and stable ways to solve such underdetermined systems via convex optimization and greedy pursuit algorithms.

To solve (1), a popular learning algorithm, known as K-SVD [6] is employed. K-SVD iteratively solves (1) by performing two steps at each iteration: (i) *sparse coding* and (ii) *dictionary update*. In the sparse coding step, $\mathbf{\Phi}$ is kept fixed and the coefficients in $\mathbf{X}$ are computed by a greedy algorithm called the *orthogonal matching pursuit* (OMP) [7]. For the details of this learning algorithm, please refer to the original work [6].

2.2. Quality Assessment

At this step, we first compare the reference and the distorted images locally. A global measure of quality is computed later by aggregating the scores obtained at the local level.

Let us consider any patch vector $\mathbf{p}_d \in \mathbb{R}^n$ extracted from the distorted image I_d and its corresponding (at the same location) patch vector $\mathbf{p}_r \in \mathbb{R}^n$ inthe reference image I_r. After obtaining the dictionary $\mathbf{\Phi}$ by solving (1), the patches $\mathbf{p}_d$ and $\mathbf{p}_r$ are decomposed using $\mathbf{\Phi}$ to obtain their respective sparse coefficient vectors $\mathbf{x}_r$ and $\mathbf{x}_d$. That is

$$\min_{\mathbf{x}_r} \quad \left\{ \|\mathbf{p}_r - \mathbf{\Phi}\mathbf{x}_r\|_2^2 \right\} \quad \text{subject to} \quad \|\mathbf{x}_r\|_0 \leq \tau \qquad (2)$$

$$\min_{\mathbf{x}_d} \quad \left\{ \|\mathbf{p}_d - \mathbf{\Phi}\mathbf{x}_d\|_2^2 \right\} \quad \text{subject to} \quad \|\mathbf{x}_d\|_0 \leq \tau \qquad (3)$$

In order to quantify the perceptual quality of $\mathbf{p}_d$ w.r.t. $\mathbf{p}_r$, we compare their sparse representations $\mathbf{x}_d$ and $\mathbf{x}_r$. A simple but effective way to compare two vectors is to compute their correlation coefficient. A scalar value α is computed based on the correlation coefficient between $\mathbf{x}_r$ and $\mathbf{x}_d$ as follows:

$$\alpha(\mathbf{p}_r, \mathbf{p}_d) = \frac{\left|\mathbf{x}_r^T \mathbf{x}_d\right| + c_1}{\|\mathbf{x}_r\|_2 \|\mathbf{x}_d\|_2 + c_1} \qquad (4)$$

where c_1 is a small positive constant added to avoid the instability caused when the denominator is close to zero. Clearly, $0 < \alpha \leq 1$. When $\mathbf{x}_r$ and $\mathbf{x}_d$ are orthogonal, we get $\left|\mathbf{x}_r^T \mathbf{x}_d\right| = 0$; but due to the presence of c_1, the parameter α is slightly greater than zero. Due to normalization, α is unaffected by the lengths of $\mathbf{x}_r$ and $\mathbf{x}_d$. Thus α is unable to measure distortions that cause the length of $\mathbf{x}_d$ to change.

To account for these types of distortions as well, we introduce another parameter. An important measure of similarity (or difference) between two vectors is their pointwise difference. Hence, we compute another quantity β which uses the length of the difference vector $(\mathbf{x}_r - \mathbf{x}_d)$.

$$\beta(\mathbf{p}_r, \mathbf{p}_d) = 1 - \frac{\|\mathbf{x}_r - \mathbf{x}_d\|_2 + c_2}{\|\mathbf{x}_r\|_2 + \|\mathbf{x}_d\|_2 + c_2} \qquad (5)$$

where c_2 is a small positive constant. It is easy to see that $0 < \beta < 1$, for non-empty $\mathbf{x}_r$ and $\mathbf{x}_d$.

It is well-known that not every pixel (or region) in an image receives the same level of importance from the perceiver. Several studies (e.g. [8, 9]) have shown that a significant improvement in the performance of quality metrics can be achieved by detecting the perceptually important regions. Hence we compute the global quality score by aggregating the local quality scores from selected regions in the image only. These regions are required to be the visually important regions of the image. Such regions are detected via an efficient saliency detection method called the spectral saliency method [10]. This is a state-of-the-art saliency detection method that analyzes the frequency spectrum of an input image obtained by the Fourier transform. The method detects the points of statistical singularities in the spectrum which corresponds to the salient regions in the spatial domain. For details, please refer to the original work [10].

Based on the saliency scores obtained using the method in [10], we select the M most salient points in I_d and their corresponding points in I_r. The global quality score $\mathcal{Q}(I_r, I_d)$ computed as follows:

$$\mathcal{Q}(I_r, I_d) = \frac{1}{M} \sum_{i=1}^{M} \alpha(\mathbf{p}_r^i, \mathbf{p}_d^i) \beta(\mathbf{p}_r^i, \mathbf{p}_d^i) \qquad (6)$$

The quality index $\mathcal{Q}$ is bounded i.e. $0 < \mathcal{Q} < 1$. This index is *not* symmetric i.e. $\mathcal{Q}(I_r, I_d) \neq \mathcal{Q}(I_d, I_r)$. This is because the dictionary $\mathbf{\Phi}$ is trained on the I_r only. Symmetry can be achieved by repeating the quality estimation stage with a dictionary trained on the distorted image and averaging the resulting quality scores obtained using the two dictionaries. Our experiments however have shown that achieving symmetry has little or no significance on the performance of the quality index.

3. PERFORMANCE EVALUATION

The proposed image quality index $\mathcal{Q}$ is evaluated on five publicly available subject-rated image quality databases - Cornell-A57 [14], CSIQ [16], LIVE [17], TID [18] and WIQ [19]. The images in these databases contain a variety of distortions including compression artifacts, blurring, flicker noise and wireless transmission artifacts.

The performance of a quality metric is evaluated by comparing the scores it yields to the scores available from human observers. To perform a comparison between the objective and subjective scores pertaining to image quality, we use a set of evaluation measures suggested by the video quality expert group (VQEG) [20] and other experts [13, 21]. These evaluation measures are - the *Spearman's rank order correlation coefficient* (SROCC), the *Pearson linear correlation coefficient* (PLCC), and *root mean squared error* (RMS).

The performance of the proposed quality metric is compared with seven well-known image quality metrics: PSNR, SSIM [1], PHVSM [11], IFC [12], VIF [13], VSNR [14], and MAD [15]. PSNR is used as a baseline method. For implementations of SSIM, PHVSM, IFC, VIF, VSNR and MAD, we have used the original MATLAB codes provided by the respective authors. The parameters of each of these methods were set to their default values as suggested in the original references.

Preprocessing: Following the preprocessing step performed in [1], all images (reference and distorted) were downsampled by a factor F so as to account for the viewing distance. The value of F was computed as
$F = \max(1, \text{round}(g/256))$
where $g = \min(\#\text{rows in } I_r, \#\text{columns in } I_r)$.

Parameter settings: In the dictionary learning step, a patch size of $\sqrt{n} \times \sqrt{n} = 11 \times 11$ was used following the patch-size specification in [1]. A collection of $k = 3000$ patches were extracted *randomly* from every reference image to train its corresponding dictionary. We set the overcompleteness factor (m/n) to 2 which yielded $m = 242$. The value of τ during dictionary learning is set to 12. The constants in (4) and (5) are chosen to have small positive values, $c_1 = 256 * 0.01$, $c_2 = 0.01$ so as they had minimal influence on the quality score. In the quality assessment step, the value of $\tau = 6$ provides the best overall trade-off between accuracy and speed. The value of M was chosen to be 15% of the total number of pixels in the distorted image.

Results: Table 1 compares the performance of the proposed quality metric with the state-of-the-art quality metrics in terms of SROCC, PLCC and RMS. A good image quality assessment metric is expected to have *high* SROCC and PLCC scores, and *low* RMS value. The best result is indicated for each database. In order to summarize the results, the average SROCC, PLCC and RMS values are computed over all datasets. The average values are computed for two cases: in the first case the values are *directly* averaged and in the second case the values are *weighted* by the size of the databases. The weight for a particular database is the number of distorted images it contains, e.g. 779 for LIVE and 54 for A57. In each case, the best two results are printed in boldface. As can be seen in Table 1, no single metric performs the best on all datasets. Our proposed metric performs the best in 3 out of the 5 datasets. On average, the proposed metric exhibits the best performance over all databases.

Computational speed: Due to its dependence on sparse coding, our proposed method is computationally demanding (but still less expensive compared to the HVS-based models like MAD). To give an idea of the computation time, a basic MATLAB implementation (on a computer with Intel Q9400 processor at 2.66 GHz) takes on average 3.4 seconds for the dictionary learning step using the parameter values specified in this paper. The quality assessment step takes about 1.0 sec.

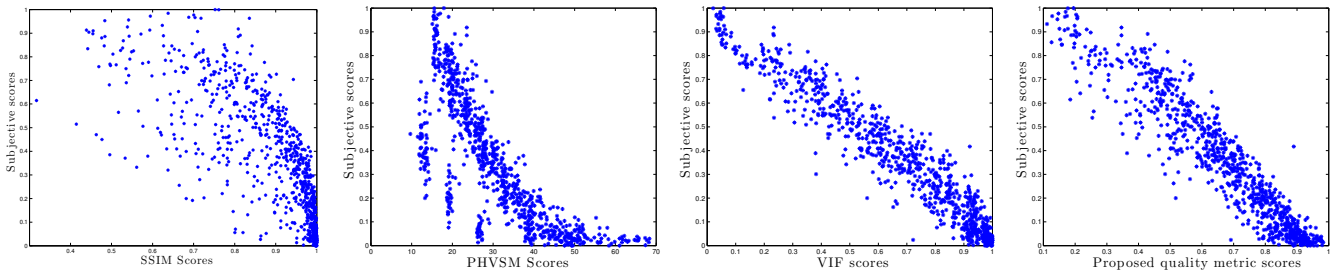

Fig. 1. Subjective scores vs. objective scores scatter plots for (left to right) SSIM, PHVSM, VIF and the proposed quality metric on the CSIQ database.

Table 1. Performance comparison of various image quality assessment metrics for five databases. Overall results (averaged over all databases) are also presented as a summary. The best result achieved in each case is written in boldface.

SROCC-based comparison								
Database	PSNR	SSIM [1]	PHVSM [11]	IFC [12]	VIF [13]	VSNR [14]	MAD [15]	Proposed
A57	0.598	0.806	0.896	0.318	0.622	**0.935**	0.864	0.920
CSIQ	0.800	0.858	0.822	0.767	0.919	0.809	0.899	**0.946**
LIVE	0.875	0.947	0.922	0.926	**0.963**	0.912	0.943	0.931
TID	0.552	0.773	0.561	0.622	0.749	0.704	0.770	**0.792**
WIQ	0.626	0.758	0.757	0.716	0.692	0.656	0.790	**0.816**
Direct avg.	0.690	0.828	0.791	0.670	0.789	0.803	0.853	**0.881**
Weighted avg.	0.688	0.833	0.716	0.723	0.836	0.779	0.843	**0.864**
PLCC-based comparison								
A57	0.628	0.802	0.875	0.372	0.614	0.914	0.881	**0.925**
CSIQ	0.746	0.758	0.772	0.821	0.927	0.735	0.820	**0.939**
LIVE	0.860	0.941	0.917	0.853	**0.944**	0.917	0.939	0.928
TID	0.519	0.727	0.552	0.660	0.808	0.682	0.748	**0.820**
WIQ	0.639	0.640	0.749	0.705	0.730	0.763	**0.830**	0.801
Direct avg.	0.678	0.774	0.773	0.682	0.804	0.802	0.843	**0.882**
Weighted avg.	0.656	0.781	0.698	0.740	0.863	0.753	0.812	**0.875**
RMS-based comparison								
A57	0.191	0.147	0.119	0.223	0.194	0.099	0.116	**0.093**
CSIQ	0.175	0.171	0.167	0.150	0.098	0.178	0.150	**0.090**
LIVE	13.990	9.985	10.892	14.263	**9.240**	10.772	9.368	10.185
TID	1.147	0.921	1.119	1.008	0.790	0.981	0.890	**0.768**
WIQ	15.426	17.595	15.185	16.252	15.653	14.809	**12.754**	13.699
Direct avg.	6.186	5.764	5.496	6.379	5.195	5.368	**4.655**	4.967
Weighted avg.	4.107	3.152	3.392	4.115	**2.857**	3.291	2.876	3.007

4. CONCLUSION

The main contribution of this work is the introduction of a novel image quality assessment metric that quantifies the perceptual quality of a distorted image with respect to a reference image. This metric compares the structural information between a distorted version of an image and its reference image based on the sparse coefficients obtained using learnt dictionary. The proposed metric performs the best in three out of the five datasets. Overall, the proposed metric is shown to outperform a number of well known image quality metrics including VIF and MAD. Our method is computationally more expensive than VIF, but is faster than MAD.

At present, the proposed quality metric works on grayscale images, requiring converting the color images to grayscale. Hence it is blind to degradations in the color components. Future work can be directed towards incorporating color information of images to improve the performance of the metric. Other possible improvements would be to learn multiscale dictionaries, and use different pooling strategies to aggregate the global score from the local quality scores.

5. ACKNOWLEDGEMENT

This work was supported by NSERC Canada and by Qatar National Research Fund (QNRF No. NPRP 09-310-1-058).

6. REFERENCES

[1] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility

to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600 –612, Apr 2004.

[2] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale structural similarity for image quality assessment," in *Asilomar Conference on Signals, Systems and Computers*, nov. 2003, vol. 2, pp. 1398 – 1402.

[3] Zhou Wang and E.P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *ICASSP*, 18-23, 2005, vol. 2, pp. 573 – 576.

[4] Guan-Hao Chen, Chun-Ling Yang, and Sheng-Li Xie, "Gradient-based structural similarity for image quality assessment," in *ICIP*, oct. 2006, pp. 2929 –2932.

[5] Zhou Wang and Xinli Shang, "Spatial pooling strategies for perceptual image quality assessment," in *ICIP 2006*, oct. 2006, pp. 2945 –2948.

[6] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, pp. 4311–4322, 2006.

[7] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proc. Asilomar Signals, Systems and Computers*, 1993.

[8] E.C. Larson and D.M. Chandler, "Unveiling relationships between regions of interest and image fidelity metrics," in *Visual Communications and Image Processing*, 2008, vol. 6822, pp. 68222A–68222A.

[9] EC Larson, C. Vu, and DM Chandler, "Can visual fixation patterns improve image fidelity assessment?," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. IEEE, 2008, pp. 2572–2575.

[10] Xiaodi Hou and Liqing Zhang, "Saliency detection: A spectral residual approach," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[11] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of dct basis functions," in *Int. Workshop Video Proc. and Quality metrics*, 2007.

[12] Hamid Rahim Sheikh, Alan Conrad Bovik, and Gustavo de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Tran. Image Processing*, vol. 14, no. 12, pp. 2117–2128, 2005.

[13] H.R. Sheikh and A.C. Bovik, "Image information and visual quality," *IEEE Tran. Image Processing*, vol. 15, no. 2, pp. 430 –444, feb. 2006.

[14] D.M. Chandler and S.S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Tran. Image Processing*, vol. 16, no. 9, pp. 2284 –2298, sep 2007.

[15] Eric C Larson and Damon M Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 011006–011006, 2010.

[16] E.C. Larson and D. M. Chandler, "Categorical image quality assessment (csiq) database," .

[17] H.R. Sheikh, Z. Wang, and A.C. Bovik, "Live image quality assessment database release 2," .

[18] N. Ponomarenko and K. Egiazarian, "Tampere image database 2008 tid2008," .

[19] U. Engelke, T.M. Kusuma, H.J. Zepernick, and M. Caldera, "Reduced-reference metric design for objective perceptual quality assessment in wireless imaging," *Signal Processing: Image Communication*, vol. 24, no. 7, pp. 525 –547, 2009.

[20] "Final report from the video quality experts group on the validation of objective models of video quality assessment," 2000.

[21] Zhou Wang and Qiang Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans Image Processing*, vol. 20, no. 5, pp. 1185 –1198, may 2011.