SEGMENTATION-BASED PERCEPTUAL IMAGE QUALITY ASSESSMENT (SPIQA)

Bernard Ghanem, Esther Resendiz, Narendra Ahuja

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
{bghanem2, eresendi, ahuja}@vision.ai.uiuc.edu

ABSTRACT

Computational representation of perceived image quality is a fundamental problem in computer vision and image processing, which has assumed increased importance with the growing role of images and video in human-computer interaction. It is well-known that the commonly used Peak Signal-to-Noise Ratio (PSNR), although analysis-friendly, falls far short of this need. We propose a perceptual image quality measure (IQM) in terms of an image's region structure. Given a reference image and its "distorted" version, we propose a "full-reference" IQM, called Segmentation-based Perceptual Image Quality Assessment (SPIQA), which quantifies this quality reduction, while minimizing the disparity between human judgment and automated prediction of image quality. One novel feature of SPIQA is that it enables the use of inter- and intra- region attributes in a way that closely resembles how the human visual system (HVS) perceives distortion. Experimental results over a number of images and distortion types demonstrate SPIQA's performance benefits.

Index Terms— IQM, QA, HVS, segmentation, saliency

1. INTRODUCTION

An IQM that creates a computational representation of perceived image quality is needed in computer vision and image processing, and has assumed increased importance with the growing role of images and video in human-computer interaction. PSNR is the de-facto standard for quality assessment due to its computational simplicity, however its use of a pixel-based distance metric fails to capture the human-perceived qualities of image distortion. Several other IQM's have been proposed recently [1], but none form a computationally robust method that mimics the HVS. Humans comprehend the contents of an image, and using mid-level techniques, an IQM should emulate the HVS by examining the structure of an image and quantifying distortion in terms of the perturbations to image structure. Segmentation is a mid-level vision technique that captures the structure of an image and achieves dimensionality reduction by dividing an image into regions that are defined by their shape, color, size, and texture. SPIQA, our proposed IQM, achieves superior results by using segmentation-based regions to quantify the distortion of an image in terms of image structure.

We compare SPIQA against PSNR and the two IQM's which had the best experimental performance in the recent survey paper [1]. Our IQM, like all two of the IQM's in [1], had superior results to PSNR. SPIQA not only outperformed the three IQM's in [1], but was able to train on only 13% of the database that was used in [1] and achieved lower RMSE with respect to human opinion scores, even before the nonlinear regression fitting that was necessary in [1].

The rest of this paper is organized as follows: **Section (2)** summarizes the previous work on IQM's, **Section (3)** gives a description of the underlying components that formulate the SPIQA measure, **Section (4)** presents experimental results of the algorithm, demonstrating its performance as compared to the IQM's that achieve the best results in [1], and **Section (5)** provides concluding comments and future goals.

2. BACKGROUND

The formulation of IQM's for image quality assessment (QA) is an old field. IQM's can be divided into two categories according to the amount and form of human intervention involved: subjective or objective IQM's. **(1)** A subjective IQM requires direct human intervention, since it is based on the cumulative judgment of a group of human observers. This type of IQM is heavily correlated with the observers' preferences.

(2) On the other hand, an objective IQM analyzes a distorted image and possibly a reference image, in the absence of any direct human intervention. Most IQM's are of this type. Objective IQM's either measure the quality of an image with respect to a reference ("full-reference") or in the absence of a reference ("blind-reference"). In this paper, we will discuss a novel, "full-reference", region-based IQM, SPIQA.

Traditional objective IQM methods rely primarily on modelling and approximating the functionality of HVS in terms of well-known image processing operations. One of the first notable IQM's was the Just Noticeable Difference (JND) measure, which was developed in the seminal work by Lubin ([2]). JND and other traditional IQM's quantify the threshold of distortion that must be exceeded before a human can perceptually detect that distortion has been imposed on the reference image. These methods tend

to fall short of efficiently approximating the complex, nonlinear functionality of the HVS. Also, some methods of this IQM type rely on parameters that are dependent on experimental settings.

More recent objective IQM's are considered signal fidelity IQM's, since they compute the measure based on inherent features of the pair of images only, thus, avoiding dependence on the experimental setup. In this work, we consider three IQM's: **(a)** the simplest and the defacto standard measure of PSNR and two signal fidelity IQM's that showed the best experimental performance in a recent survey [1]: **(b)** Multi-Scale Structural SIMilarity (MSSSIM) [3] and **(c)** Visual Information Fidelity (VIF) [4].

(a) PSNR uses a pixel-based distance measure. However, this method fails to capture the structure of distortions. Such structure plays an important role in perception of distortion by humans, and occurs in most applications (e.g. blocking artifacts due to JPEG compression).

(b) MSSSIM divides an image into rectangular blocks, or patches, and computes 1^{st} and 2^{nd} order statistics for each patch. These statistics do not sufficiently represent the luminance distribution of image patches. Also, its "structural factor" is independent of the spatial distributions. The unexplored spatial relationships are important, as evident from the *visual masking* phenomenon [5], where respective luminance distributions and spatial localization of the image regions are able to mask or enhance distortion in a specific region.

(c) VIF takes an information fidelity approach to image QA using wavelet decomposition. No explicit interaction between wavelet subbands is modeled, since the subband-specific VIF measures are pooled together independently. This independence counteracts the highly regarded *contrast sensitivity function* (CSF) [5], which renders certain wavelet subbands less effective than others in quality perception.

In this paper, we propose a signal fidelity IQM (SPIQA), which manipulates some of the important characteristics of the HVS, while remaining independent of subjective factors related to the experimental setup. The contributions of our proposed IQM are three fold: **(1)** it uses image segmentation to delineate coherent regions of human attention, **(2)** it quantifies both inter- and intra- region interactions in a manner that conforms to certain functional aspects of the HVS, and **(3)** it quantifies the quality of an image segment via local (e.g. at ramp and non-ramp pixels) and global features that represent both the structure and the content of each segment.

3. REGION BASED IMAGE QUALITY ASSESSMENT

In this section, we describe how SPIQA is formulated. The major motivation for our measure is to incorporate image segments in its definition, which makes the quality measure depend on spatial structure in addition to image intensity values.

Image segmentation partitions an image into disjoint regions that contain pixels that are "similar" to each other, but "different" from the pixels of another region. The problem of efficient and perceptually correct segmentation is still an unsolved problem in computer vision, but there are numerous algorithms in the literature that approximate the segmentation. We use the multi-scale segmentation algorithm implemented in [6]. The resulting segmentation is characterized by homogeneous regions surrounded by ramp discontinuities. Thus, each segment at every photometric scale includes ramp and non-ramp pixels. In our implementation, we use a single photometric scale and segment only the reference image. The same segment boundaries are also used for the distorted image.

After image segmentation, we compute the overall measure as a weighted sum of the regional image quality measures (RIQM's). Each segment's RIQM contributes to SPIQA as follows:

$$\text{SPIQA} = \sum_{\mathbf{seg}_i \subset \mathbf{I}_{\text{ref}}} w_i \ \text{RIQM}_i \tag{1}$$

$$w_i = \beta_1 \text{size}_i + (1-\beta_1)\text{sal}_i \ ; \quad \beta_1 \in [0,1] \tag{2}$$

where w_i weighs the contribution of the RIQM in the i^{th} segment, thus summarizing all inter-segment interactions by quantifying the importance of each segment in terms of its overall saliency and size. RIQM_i is the regional quality measure which summarizes all intra-segment interactions according HVS perception of independent segments.

3.1. Inter-Segment Interactions: w_i

Equation (2) expresses w_i as a linear combination of two normalized factors: $\text{size}_i = \frac{\#\text{ of pixels in seg}_i}{\#\text{ of pixels in } \mathbf{I}_{ref}}$ and $\text{sal}_i = \frac{\text{saliency of seg}_i}{\text{saliency of } \mathbf{I}_{ref}}$. Here, saliency is computed in accordance to human visual attention as described in [7] (refer to Figure 1). We justify the use of saliency from a human perception point of view. Humans concentrate on high-level features of an image to identify its contents; however the saliency algorithm computes the saliency map on a pixel basis using low-level features.

By incorporating the pixel-based saliency map into coherent regions, we incorporate higher-level features that better represent the focus of human attention. By virtue of w_i's, the effects of distortion on the image quality are more influenced by the distortions in the most salient regions. To the best of our knowledge, the use of such segment interaction is novel to non-traditional IQM formulation, which usually assumes independence between neighboring blocks or bands.

3.2. Intra-Segment Interactions: RIQM_i

This section highlights the intra-segment interactions, which capture the "similarity" between the reference and distorted segment. The YCrCb color space (primarily the luminance component) is used, since it best approximates the HVS color perception among common color spaces. RIQM is defined as the product of three factors, as shown in Equation (3): namely gradient similarity (ΔG_i), similarity in histogram (ΔH_i), and normalized mutual

Fig. 1. The reference image is on the left and its corresponding regional saliency map is on the right. Lighter regions indicate higher saliency.

information(ΔNM_i). All three factors are properly normalized to take on values in the range [0,1].

$$RIQM_i = (\Delta G_i)^{\beta_3}(\Delta H_i)^{\beta_4}(\Delta NM_i)^{\beta_5} \quad (3)$$

Gradient Similarity (Δ**G**): We quantify the difference in Sobel gradient energy for each reference and distorted region. This structural term is absent in PSNR, MSSSIM, and VIF. Based on [6], a segment contains either significant ramp or non-ramp pixels, which are distinguished according to the variations of their luminance values with their neighbors. We evaluate the perceptual effects of distortion on ramp and non-ramp pixels separately. Therefore, we can write $\Delta\mathbf{G}$ of a segment (p) as a linear combination of two terms: the gradient similarity at significant ramp pixels ($\Delta\mathbf{G}_{p_r}$) and the gradient similarity at non-significant ramp pixels ($\Delta\mathbf{G}_{p_{nr}}$) as in Equation (4) with $\beta_2 \in [0,1]$.

$$\Delta G_p = \beta_2 \Delta G_{p_r} + (1-\beta_2)\Delta G_{p_{nr}} \quad (4)$$

where $\Delta\mathbf{G}_{p_r}$ and $\Delta\mathbf{G}_{p_{nr}}$ are defined as

$$\Delta G_{p_r} = \frac{2\overrightarrow{G}_{\text{ref}_{p_r}} \cdot \overrightarrow{G}_{\text{dis}_{p_r}}}{\|\overrightarrow{G}_{\text{ref}_{p_r}}\|^2 + \|\overrightarrow{G}_{\text{dis}_{p_r}}\|^2 + \epsilon} \quad (5)$$

$$\Delta G_{p_{nr}} = \frac{2\overrightarrow{G}_{\text{ref}_{p_{nr}}} \cdot \overrightarrow{G}_{\text{dis}_{p_{nr}}}}{\|\overrightarrow{G}_{\text{ref}_{p_{nr}}}\|^2 + \|\overrightarrow{G}_{\text{dis}_{p_{nr}}}\|^2 + \epsilon} \quad (6)$$

The Sobel gradient energy vector, $\overrightarrow{G}_x$, that is used in (4)-(6) is computed from the gradient of the image I: $G_x(i) = \|\nabla I_x(i)\|$, where x defines the specific region components ($p_r|p_{nr}$) and image type (dis|ref) and i defines a pixel in x.

Histogram Similarity (Δ**H**): This term is a non structural factor that measures the difference in the distribution (estimated by a histogram) of the luminance values of the pixels within the reference and distorted segment. We define it as: $\Delta H_p = \frac{2\overrightarrow{H}_{\text{ref}_p} \cdot \overrightarrow{H}_{\text{dis}_p}}{\|\overrightarrow{H}_{\text{ref}_p}\|^2 + \|\overrightarrow{H}_{\text{dis}_p}\|^2 + \epsilon}$. This factor improves on the SSIM measure, since the difference in luminance distributions encompasses more information than simply the 1st and 2nd moments.

Normalized Mutual Information Similarity (Δ**NM**): This is another non-structural factor, which builds on the assumption made in [4] that the HVS reacts to the loss in mutual information between the reference and distorted images. It normalizes the segment's mutual information by the entropy of the reference segment. We define it as: $\Delta NM_p = \frac{I(I_{\text{ref}_p}; I_{\text{dis}_p})}{H(I_{\text{ref}_p})}$.

We determine all five β values by minimizing the squared error between the resultant SPIQA and the experimental human decisions, in difference mean opinion score (DMOS) format used in [1], over a set of N training image pairs as shown below.

$$\overrightarrow{\beta}^* = \arg\min \sum_{t=1}^{N} |\text{DMOS}(t) - \text{SPIQA}(t, \overrightarrow{\beta})|^2$$

4. EXPERIMENTAL RESULTS

We evaluate our IQM on the LIVE database used by [1], which presents the most recent and comprehensive survey of the performance of various IQM's available in literature. We will compare SPIQA with the IQM's examined in [1] (i.e. MSSSIM and VIF) using the experimental human results that [1] presents in normalized DMOS format. The LIVE database contains 29 reference images that are distorted by one of the five distortion types: JPEG2000 (227 images), JPEG (233 images), white noise - WN (174 images), Gaussian blur - GBlur (174 images), and fast fading - FF (174 images).

First, we show empirical evidence that justifies our use of segments instead of regular rectangular blocks. We applied the SSIM algorithm to both segments and blocks. Table 1 shows the improvement in RMSE performance of a segment-based SSIM over the block-based SSIM.

	RMSE
Block SSIM	18.75
Segment SSIM	6.120

Table 1. Block SSIM vs. Segment SSIM

Next, we learn the β values from 13% of the image pairs in the database (i.e. 20 from each of the five distortion types or $N = 100$ image pairs) and compare the performance of SPIQA against that of MSSSIM and VIF on the entire database. In Figure 2, we plot the ground truth human judgment values (DMOS) and the normalized IQM's generated by MSSSIM, VIF, and our proposed method when applied to each database image. The more the IQM curve approximates the ground truth, the closer the IQM is to human judgment. The top and bottom plots show the results before and after nonlinear regression (as described in [1]) respectively. For visual purposes, we only consider a portion of the LIVE database in these plots. The impact of nonlinear regression on both VIF and MSSSIM is quite significant, while it is incremental for SPIQA.

Table 2 summarizes the performance of SPIQA, VIF, MSSSIM, and PSNR. In [1], all the database samples are used for training, and in Table 2 we perform our experiments in the same fashion. These experiments show that SPIQA outperforms other IQM's, despite nonlinear regression.

	PSNR	MSSSIM	VIF	SPIQA
JPEG2000	10.61	5.999	5.093	5.076
JPEG	12.17	5.465	5.318	5.585
WN	4.669	6.358	4.360	3.920
GBlur	11.44	5.823	3.991	4.117
FF	12.97	10.40	6.855	3.519
All Data	13.43	9.369	8.246	6.546

Table 2. RMSE Comparison - SPIQA and other IQM's

Table 3 shows the numerical values for the estimated β values. Also, by using the database images of each distortion type in isolation, we determined the optimal β values for each distortion type. From these results, we can make the following key observations. For β_1, regional saliency is the single inter-regional factor to be maintained. This term inherently depends on segment size, as it is the normalized sum of all saliency values within a segment. For β_2, significant ramp pixels are more effective in detecting change in image quality, especially for distortion types that impose structured alteration close to strong edges (e.g. JPEG2000). For β_3, ΔG plays the most influential role in QA. This is due to the fundamental impact of structured organization on human visual perception. For β_4, ΔH is critical in evaluating distortion types that produce significant disruptions in regional luminance distribution (e.g. Gaussian blur). But, the HVS seems to tolerate more change in ΔH than ΔG. For β_5, ΔNM is the least important factor, despite its informational description of human visual judgment.

	β_1	β_2	β_3	β_4	β_5
All Types	0	0.872	2.407	0.670	0.255

Table 3. SPIQA weights based on $N = 100$ samples

5. CONCLUSION AND FUTURE WORK

In this paper, we present a novel segmentation-based IQM, which models both inter- and intra-segment relationships, thus, capturing the HVS characteristics more effectively than previous IQM's. SPIQA improves over the state-of-the-art quality measures by reducing the gap between automatic prediction and human judgment of image quality.

For future work, we propose to conduct more extensive human experiments, which highlight how the HVS reacts to a mixture of distortion types in the same image. We also will experiment with various photometric scales in the segmentation algorithm to evaluate an optimal scale, and later incorporate this into a multi-scale algorithm.

6. ACKNOWLEDGEMENT

This research was supported by the Office of Naval Research under grant: N00014-06-1-0101.

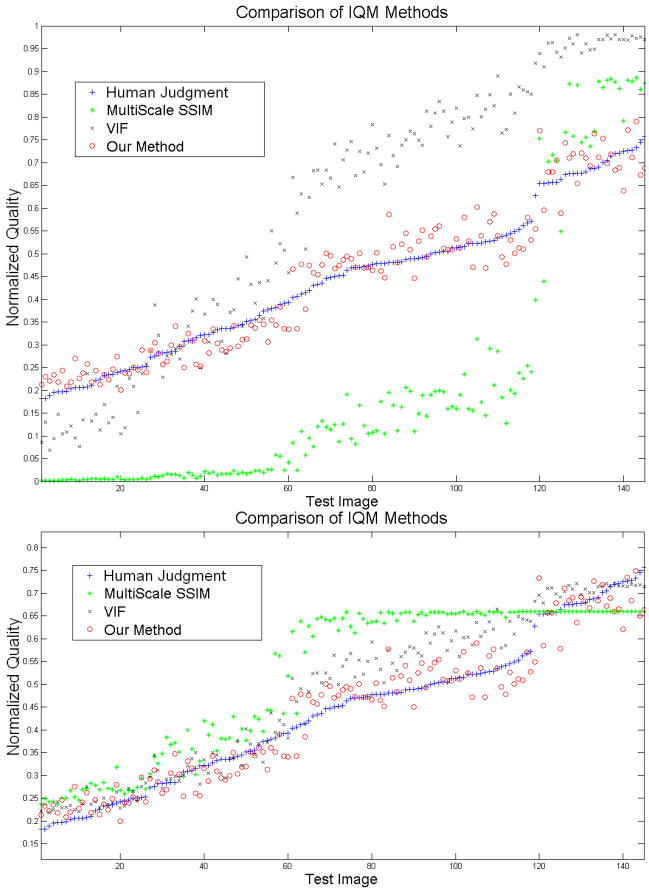

Fig. 2. SPIQA, VIF, and MSSSIM, before (top) and after (bottom) regression

7. REFERENCES

[1] H. R. Sheikh, M. R. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. on Image Processing*, vol. 15, Nov. 2006.

[2] J. Lubin, *Visual Models for Target Detection and Recognition*, 2^{nd} edition, 1995.

[3] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*, Nov. 2003.

[4] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. on Image Processing*, vol. 15, Feb. 2006.

[5] A. N. Netravali and B. G. Haskell, *Digital Pictures*, 1995.

[6] H. Arora and N. Ahuja, "Analysis of ramp discontinuity model for multiscale image segmentation," in *Proc. of ICPR*, Aug. 2006, vol. 4, pp. 99–103.

[7] L. Itti, C. Koch, and E. Neibur, "A model of saliency-based visual attention for rapid scene analysis," *PAMI*, vol. 20, pp. 1254–1259, 1998.