SPATIAL POOLING STRATEGIES FOR PERCEPTUAL IMAGE QUALITY ASSESSMENT

Zhou Wang and Xinli Shang

Dept. of EE, Univ. of Texas at Arlington, Arlington, TX 76019
Email: zhouwang@ieee.org

ABSTRACT

Many recently proposed perceptual image quality assessment algorithms are implemented in two stages. In the first stage, image quality is evaluated within local regions. This results in a quality/distortion map over the image space. In the second stage, a spatial pooling algorithm is employed that combines the quality/distortion map into a single quality score. While great effort has been devoted to developing algorithms for the first stage, little has been done to find the best strategies for the second stage (and simple spatial average is often used). In this work, we investigate three spatial pooling methods for the second stage: Minkowski pooling, local quality/distortion-weighted pooling, and information content-weighted pooling. Extensive experiments with the LIVE database show that all three methods may improve the prediction performance of perceptual image quality measures, but the third method demonstrates the best potential to be a general and robust method that leads to consistent improvement over a wide range of image distortion types.

Index Terms: image quality assessment, visual perception, structural similarity, error pooling, information content

1. INTRODUCTION

Recently, a number of objective image quality assessment algorithms have been proposed to predict human perception of image quality [1]. They may be classified into full-reference algorithms (where an original "perfect-quality" image is available as a reference), reduced-reference algorithms (where only partial information about the original image is accessible) and no-reference algorithms (where no information about the original image is available). Many of these algorithms (especially full-reference algorithms) adopted a two-stage implementation: In the first stage, image quality/distortion is evaluated locally within small regions, resulting in a quality/distortion map. In the second stage, a spatial pooling algorithm is employed to combine the quality/distortion map into a single quality score. Such a two-stage approach may be applied directly in image pixel domain or after channel decompositions (e.g., applied to a wavelet subband).

A pixel-domain full-reference example is shown in Fig. 1, where the goal is to evaluate the quality of Image (b) with a given perfect-quality reference Image (a). Two methods are used to compute local quality/distortions – absolute difference and the structural similarity (SSIM) index [2]. The resulting quality/distortion maps are shown in Figs. 1(c) and 1(d), respectively. For easy comparison, we have adjusted the quality/distortion map representations so that brighter indicates better quality in both maps. Careful inspection shows that the SSIM index (computed within a local window that slides across the image space) better reflects the spatial variations of perceived image quality. For example, the blockiness in the sky is clearly indicated in Fig. 1(d) but not in Fig. 1(c). However, the major concern here is not on how to create a better quality map but on how to convert a quality map into a scalar quality score.

Surprisingly, in the literature, little investigation and careful comparison have been devoted to developing and testing spatial pooling methods. In practice, spatial pooling has often been treated superficially, e.g., using a simple spatial average. Some methods incorporate human interactions or automatic object detections and segmentations to define the regions-of-interest or points-of-fixations before spatial pooling (e.g., [3,4]), but these methods may not be easily applied to general-purpose image quality assessment because for many images, it may not always be easy to find obviously outstanding objects that attract visual attention. On the other hand, problems arise with the direct spatial average approach when the distortion is highly non-uniform over the image space. For example, when only a small region in an image is corrupted with extremely annoying artifacts, but all other regions have high quality, human subjects tend to pay more attention to the low quality region and give an overall quality score lower than the average of the quality/distortion map.

In this paper, we will have a close look at three well-motivated strategies for spatial pooling – Minkowski pooling, local quality/distortion weighted pooling, and information content-weighted pooling. The questions that we would like to answer are: 1) do these spatial pooling strategies, as compared to simple spatial average, improve the prediction capability of perceptual image quality measures? 2) Is such improvement consistent for different types of quality/distortion maps and over a wide variety of image distortion types?

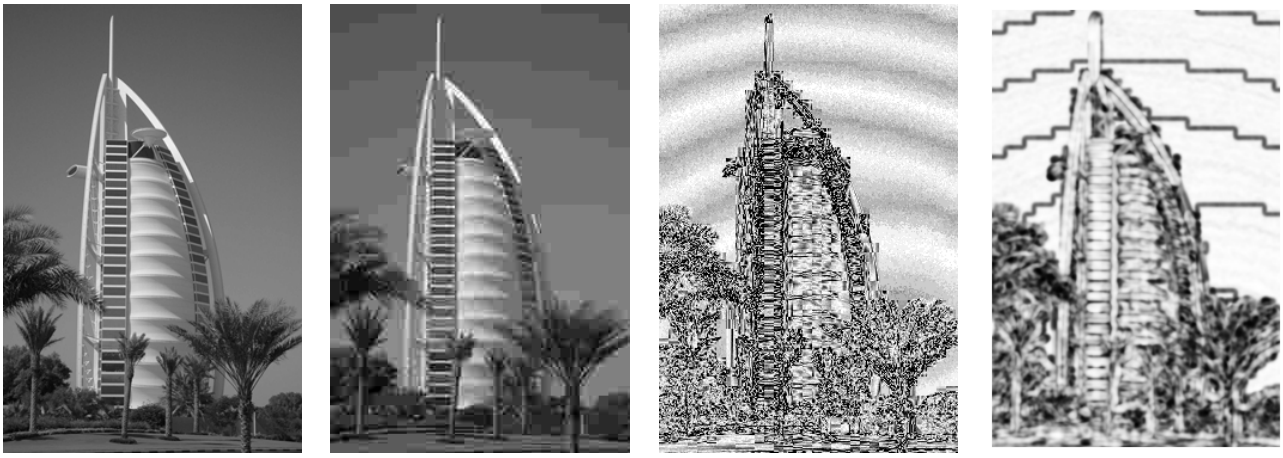
2. SPATIAL POOLING STRATEGIES

2.1. Minkowski Pooling

Let m_i be the quality/distortion value at the i-th spatial location in the quality/distortion map. The non-uniform quality distribution problem discussed in Section 1 may be partially solved (though in an indirect way) by adopting a Minkowski pooling approach, which has been extensively used [1] and is defined as

$$M = \frac{1}{N}\sum_{i=1}^{N} m_i^p \,, \tag{1}$$

where N is the number of samples in the quality/distortion map, and p is the Minkowski power. As a special case, when m_i represents the absolute difference as in Fig. 1(c), then Eq. (1) is directly related to the l_p norm (subject to a normalization constant and a



(a) (b) (c) (d)

Fig. 1. (a) Original image; (b) distorted image (created by JPEG compression); (c) absolute difference map: brighter indicates better quality (smaller absolute difference between the original and the distorted images); (d) SSIM index map: brighter indicates better quality (larger local SSIM value).

monotonic nonlinearity). In particular, when p = 1, it reduces to the mean absolute error (MAE). When p = 2, it becomes the mean squared error (MSE), which can be monotonically mapped to the widely used peak signal-to-noise ratio (PSNR). As p increases, more emphasis will be put at the image regions of high distortions. It is often conjectured that an appropriate value of p should provide a reasonable estimation of how humans rate image quality.

2.2. Local Quality/Distortion-Weighted Pooling

The non-uniform quality distribution problem may also be solved more directly by assigning spatially varying importance (weights) over the image space. A general form of such a spatial weighting approach is given by

$$M = \frac{\sum_{i=1}^{N} w_i m_i}{\sum_{i=1}^{N} w_i}\,, \tag{2}$$

where w_i is the weight assign to the i-th spatial location.

The idea of local quality/distortion-weighted pooling is to define the weight w_i by the local quality measure m_i itself, i.e.,

$$w_i = f(m_i)\,. \tag{3}$$

For example, in the case that m_i represents a distortion measure (higher value indicates higher distortion) and we would like to put more emphasis on the spatial locations where the image quality is extremely bad, then we would choose $f(\cdot)$ to be a monotonically increasing function. On the other hand, if m_i is a quality measure (higher value indicates better quality), then we would prefer $f(\cdot)$ to be a monotonically decreasing function.

2.3. Information Content-Weighted Pooling

In information content-weighted pooling, a similar spatial weighting method as in Eq. (2) is employed. However, the weights are determined by the local image content (of either or both of the reference and the distorted images), rather than the measured local quality/distortion. Let $\mathbf{x}_i$ and $\mathbf{y}_i$ be the local image patches (e.g., a collection of pixels in a local window) extracted around the i-th spatial location from the reference and the distorted images, respectively. The weight w_i is computed using a function

$$w_i = g(\mathbf{x}_i, \mathbf{y}_i)\,. \tag{4}$$

The local energy-weighted pooling method proposed in [5] may be considered as a special case of this apporach, where the weighting function is given by

$$g(\mathbf{x}, \mathbf{y}) = \sigma_x^2 + \sigma_y^2 + C\,. \tag{5}$$

Here σ_x and σ_y are the standard deviations of $\mathbf{x}$ and $\mathbf{y}$, respectively, and C is a constant representing a baseline minimal weight. The underlying justification of using Eq. (5) is that the high-energy (or high-variance) image regions are likely to contain more information. If the ultimate goal of visual perception is to efficiently extract useful information from the visual scene, then the high-energy regions are more likely to attract visual attention, and thus should be given more importance. While this general idea is well-motivated, the specific formulation of Eq. (5) is not directly an information measure based on any statistical model.

Here we propose a new method, in which the perceived local information content is quantified as the number of bits that can be received from a statistical image information source that passes through a noisy visual channel. To keep the algorithm tractable, we assume a local Gaussian source model and an additive Gaussian channel model. Similar information communication-based models had been used previously for image quality assessment [6], though not involved in the spatial pooling stage. Assume that the source power is S and the channel noise power is C (which is considered as an estimate of the intrinsic noise in the visual system [6]). A well-known result from information theory is that the received information can be computed as

$$I = \frac{1}{2}\log\left(1 + \frac{S}{C}\right)\,. \tag{6}$$

Now assume that the source power of a local image patch $\mathbf{x}$ can be estimated as σ_x^2, and the channel noise variance is a known

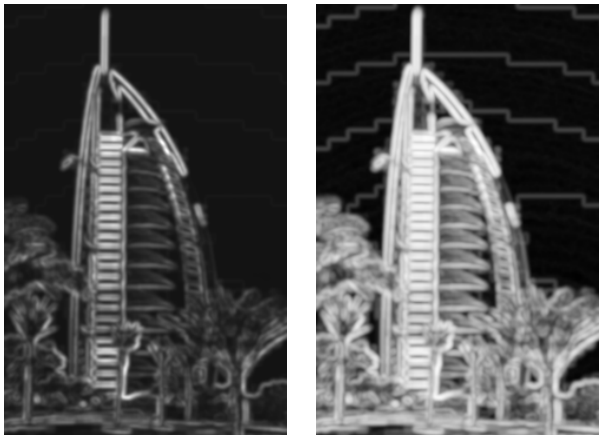
Table 1. Performance comparison of spatial pooling methods. The absolute difference is used to generate the distortion map. JP2: JPEG2000 dataset; JPG: JPEG; Noise: white Gaussian noise; Blur: Gaussian blur; Error: transmission error; AI: average improvement.

method			LIVE dataset / ROCC result							
pooling strategy	p	w_i	JP2-1	JP2-2	JPG-1	JPG-2	Noise	Blur	Error	AI
spatial average	1	1	0.9026	0.8180	0.8722	0.7485	0.9857	0.7425	0.8651	0
Minkowski pooling	1/8	1	0.8619∨	0.7384∨	0.8562∨	0.7487−	0.9860−	0.6309∨	0.7669∨	−0.0494
	1/4	1	0.8700∨	0.7595∨	0.8593∨	0.7541∧	0.9860−	0.6621∨	0.7998∨	−0.0348
	1/2	1	0.8823∨	0.7875∨	0.8607∨	0.7478−	0.9858−	0.7003∨	0.8386∨	−0.0188
	2	1	0.9227∧	0.8662∧	0.8876∧	0.7446∨	0.9856−	0.7921∧	0.8931∧	+0.0225
	4	1	0.9449∧	0.9105∧	0.9052∧	0.7573∧	0.9845−	0.8413∧	0.9012∧	+0.0443
	8	1	0.9566∧	0.9438∧	0.9355∧	0.7934∧	0.9843−	0.8731∧	0.8931∧	+0.0636
local quality /distortion-weighted pooling	1	$\lvert m_i\rvert^{1/8}$	0.9152∧	0.8431∧	0.8721−	0.7479−	0.9855−	0.7536∧	0.8840∧	+0.0095
	1	$\lvert m_i\rvert^{1/4}$	0.9204∧	0.8539∧	0.8753∧	0.7438∨	0.9853−	0.7671∧	0.8875∧	+0.0141
	1	$\lvert m_i\rvert^{1/2}$	0.9280∧	0.8709∧	0.8858∧	0.7385∨	0.9849−	0.7856∧	0.8956∧	+0.0221
	1	$\lvert m_i\rvert^{1}$	0.9412∧	0.8956∧	0.8944∧	0.7359∨	0.9844−	0.8218∧	0.9006∧	+0.0342
	1	$\lvert m_i\rvert^{2}$	0.9529∧	0.9302∧	0.9173∧	0.7457−	0.9841−	0.8470∧	0.8873∧	+0.0471
	1	$\lvert m_i\rvert^{4}$	0.9592∧	0.9485∧	0.9360∧	0.8068∧	0.9836−	0.8514∧	0.8550∨	+0.0580
	1	$\lvert m_i\rvert^{8}$	0.9594∧	0.9461∧	0.9487∧	0.8453∧	0.9826∨	0.8412∧	0.8466∨	+0.0622
info. content-weighted pooling	1	Eq. (5)	0.9512∧	0.9341∧	0.9294∧	0.7850∧	0.9858−	0.8287∧	0.9214∧	+0.0573
	1	Eq. (7)	0.9556∧	0.9332∧	0.9210∧	0.7864∧	0.9859−	0.8809∧	0.9299∧	**+0.0655**

parameter (as in [6]). Then the weighting function is given by

$$g(\mathbf{x},\mathbf{y}) = \log\left[\left(1+\frac{\sigma_x^2}{C}\right)\left(1+\frac{\sigma_y^2}{C}\right)\right]. \qquad (7)$$

Here we have removed the front scalar constant, which has no effect on the final pooling result because of the normalization in Eq. (2). We have also added the information content of both the reference and the distorted image patches, so as to make the algorithm symmetric. Figure 2 gives an example of an information content-based weighting function over the image space, which is computed for the images shown in Fig. 1. As in [2], in the computation of local σ_x^2 and σ_y^2, a sliding Gaussian window with standard deviation of 1.5 pixels is employed.



(a) (b)

Fig. 2. Weighting function calculated based on local information content of Images (a) and (b) in Fig. 1. (a) computed using Eq. (5); (b) computed using Eq. (7).

3. TEST

We test the objective image quality measures with different spatial pooling approaches using the LIVE database [7], which contains seven subject-rated datasets, including two datasets for JPEG 2000 compression (contains 87 and 82 images, respectively), two for JPEG compression (contains 87 and 88 images, respectively), one for white Gaussian noise contamination (145 images), one for Gaussian blur (145 images), and one for transmission errors of JPEG 2000 compressed images (145 images). For each objective quality measure being evaluated, we report the Spearman rank-order correlation coefficients (ROCC) between the subjective and objective scores for each dataset. The ROCC is defined as

$$r = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{K(K^2-1)}, \qquad (8)$$

where K is the number of images in the dataset, and d_i is the difference between the i-th image's ranks in subjective and objective evaluations. ROCC is one of the metrics adopted by the video quality experts group (VQEG) for the evaluation of video quality measures [8]. Similar results are obtained by other VQEG metrics, though not reported here because of the space limit.

The image quality measures being evaluated are divided into two groups. The first group uses the absolute difference to create the distortion map, and the second group uses the SSIM index to generate the quality map (as in [2], the SSIM index maps are computed after downsampling the images by a factor of 2). For each group, we first use the simple spatial average as the pooling method. The results will then be used as the baseline performance to compare the other pooling methods. Next, six Minkowski pooling methods are tested, where the values of p are 1/8, 1/4, 1/2, 2, 4, and 8, respectively. Seven local quality/distortion-weighted pooling methods are also tested for each group, where we use $f(m_i) = |m_i|^q$ to compute the weighting functions and the values of q are 1/8, 1/4, 1/2, 1, 2, 4, and 8 for the first group and $-1/8$, $-1/4$, $-1/2$, -1, -2, -4, and -8 for the second group, respectively.

Table 2. Performance comparison of spatial pooling methods. The SSIM index is used to generate the quality map. JP2: JPEG2000 dataset; JPG: JPEG; Noise: white Gaussian noise; Blur: Gaussian blur; Error: transmission error; AI: average improvement.

method			LIVE dataset / ROCC result							
pooling strategy	p	w_i	JP2-1	JP2-2	JPG-1	JPG-2	Noise	Blur	Error	AI
spatial average	1	1	0.9545	0.9636	0.9598	0.9028	0.9737	0.9497	0.9546	0
Minkowski pooling	1/8	1	0.9549−	0.9660−	0.9609−	0.9069∧	0.9777∧	0.9559∧	0.9554−	+0.0027
	1/4	1	0.9547−	0.9652−	0.9608−	0.9063∧	0.9768∧	0.9552∧	0.9554−	+0.0022
	1/2	1	0.9542−	0.9642−	0.9605−	0.9035−	0.9755−	0.9531∧	0.9551−	+0.0011
	2	1	0.9537−	0.9620−	0.9589−	0.8978∨	0.9712−	0.9430∨	0.9529−	−0.0027
	4	1	0.9506∨	0.9551∨	0.9573−	0.8808∨	0.9707∨	0.9321∨	0.9505∨	−0.0088
	8	1	0.9473∨	0.9443∨	0.9556∨	0.8662∨	0.9712−	0.9078∨	0.9447∨	−0.0174
local quality /distortion-weighted pooling	1	$\lvert m_i\rvert^{-1/8}$	0.9541−	0.9637−	0.9600−	0.9023−	0.9743−	0.9513−	0.9550−	+0.0003
	1	$\lvert m_i\rvert^{-1/4}$	0.9544−	0.9642−	0.9605−	0.9030−	0.9755−	0.9537∧	0.9552−	+0.0011
	1	$\lvert m_i\rvert^{-1/2}$	0.9552−	0.9661−	0.9609−	0.9083∧	0.9779∧	0.9566∧	0.9551−	+0.0031
	1	$\lvert m_i\rvert^{-1}$	0.9577∧	0.9698∧	0.9606−	0.9114∧	0.9825∧	0.9603∧	0.9492∨	+0.0047
	1	$\lvert m_i\rvert^{-2}$	0.9617∧	0.9708∧	0.9613−	0.9096∧	0.9849∧	0.9640∧	0.9381∨	+0.0045
	1	$\lvert m_i\rvert^{-4}$	0.9638∧	0.9678∧	0.9627−	0.8527∨	0.9592∨	0.9603∧	0.8797∨	−0.0161
	1	$\lvert m_i\rvert^{-8}$	0.9673∧	0.9615−	0.9629∧	0.8664∨	0.9584∨	0.9507−	0.8668∨	−0.0178
info. content-weighted pooling	1	Eq. (5)	0.9535−	0.9671∧	0.9439∨	0.9288∧	0.9723−	0.9672∧	0.9662∧	+0.0058
	1	Eq. (7)	0.9612∧	0.9743∧	0.9591−	0.9401∧	0.9776∧	0.9716∧	0.9659∧	**+0.0130**

Finally, two information content-weighted pooling methods are tested, in which the weighting functions are computed by Eq. (5) and Eq. (7), respectively.

The ROCC results for the two groups of objective image quality measures are shown in Table 1 and Table 2, respectively. For easy visualization, we have added a "∧", a "∨" or a "−" mark behind each ROCC number to indicate an increase/decrease/no-significant-change of ROCC value as compared to the baseline ROCC (given by spatial average pooling). We have also added a final column that gives the average improvement of ROCC values over the baseline. It can be observed that all three pooling strategies may lead to improvement of quality prediction performance. However, the best parameter choices of the Minkowski pooling methods and the local quality/distortion-weighted pooling methods depend on the underlying specific local quality/distortion measure. For example, Minkowski pooling with p = 4 results in improvement when the local quality/distortion measure is the absolute difference, but not the SSIM index. Comparatively, the information content-weighted pooling method, especially when the newly proposed Eq. (7) is used as the weighting function, appears to be more stable and general. It results in consistent and most of the time significant improvement over a wide range of image distortion types for both cases of local quality/distortion measures.

4. CONCLUSION

We have tested three spatial pooling strategies for perceptual image quality assessment based on an extensive experiment with the LIVE database. Our results suggest that all three methods may improve the prediction performance of image quality measures. Among them, the newly proposed information content-weighted pooling approach demonstrates the best potential to be a general and stable approach that provides consistent improvement over a wide range of image distortion types. Future work includes investigating the dependencies between the pooling strategies and the local quality/distortion measures, testing the pooling methods with other local quality/distortion measures (including those that involve wavelet decompositions), and developing improved method for the estimation of local information content by adopting advanced statistical image models.

5. REFERENCES

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

[2] Z. 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, pp. 600–612, Apr. 2004.

[3] W. Osberger, N. Bergmann, and A. Maeder, "An automatic image quality assessment technique incorporating high level perceptual factors," in *Proc. IEEE Int. Conf. Image Proc.*, pp. 414–418, 1998.

[4] Z. Wang, L. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Processing*, vol. 12, pp. 243–254, Feb. 2003.

[5] Z. Wang and E. P. Simoncelli, "Stimulus synthesis for efficient evaluation and refinement of perceptual image quality metrics," in *Human Vision and Electronic Imaging IX, Proc. SPIE*, vol. 5292, pp. 99–108, Jan. 2004.

[6] 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 Processing*, vol. 14, pp. 2117–2128, Dec. 2005.

[7] H. R. Sheikh, Z. Wang, A. C. Bovik, and L. K. Cormack, "Image and video quality assessment research at LIVE," `http://live.ece.utexas.edu/research/quality/`.

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