IEEE International Conference on Image Processing (ICIP12), Orlando, FL, Sept. 2012

OBJECTIVE QUALITY ASSESSMENT FOR IMAGE SUPER-RESOLUTION: A NATURAL SCENE STATISTICS APPROACH

Hojatollah Yeganeh, Mohammad Rostami and Zhou Wang

Dept. of Electrical & Computer Engineering, University of Waterloo, Waterloo, ON, Canada
Email: hyeganeh@uwaterloo.ca, m2rostami@uwaterloo.ca, zhouwang@ieee.org

ABSTRACT

There has been an increasing number of image super-resolution (SR) algorithms proposed recently to create images with higher spatial resolution from low-resolution (LR) images. Nevertheless, how to evaluate the performance of such SR and interpolation algorithms remains an open problem. Subjective assessment methods are useful and reliable, but are expensive, time-consuming, and difficult to be embedded into the design and optimization procedures of SR and interpolation algorithms. Here we make one of the first attempts to develop an objective quality assessment method of a given resolution-enhanced image using the available LR image as a reference. Our algorithm follows the philosophy behind the natural scene statistics (NSS) approach. Specifically, we build statistical models of frequency energy falloff and spatial continuity based on high quality natural images and use the departures from such models to quantify image quality degradations. Subjective experiments have been carried out that verify the effectiveness of the proposed approach.

Index Terms— image quality assessment, image super-resolution, image interpolation, natural scene statistics

1. INTRODUCTION

Image super-resolution (SR) techniques improve the spatial resolution of images beyond the limitations of the imaging acquisition devices. These techniques play important roles in a variety of applications such as web browsing, medical imaging, and high-definition television (HDTV) [1]. Here we are interested in SR algorithms that use a single low-resolution (LR) image as the input and generates a high-resolution (HR) image. In this case, image interpolation methods may be applied, where the LR image is assumed to be a directly downsampled version of the HR image where the pixel intensities remain unchanged at the sampling points. However, generally SR approaches may not strictly follow this assumption, and may alter the intensity values of the sampling pixels.

Although an increasing number of SR and interpolation algorithms have been proposed in recent years, how to evaluate their performance remains an open problem [2, 3]. A straightforward approach is subjective evaluation [3, 4], where multiple subjects are asked to rate the quality of resolution-enhanced images and the mean opinion scores (MOS) of the subjects is used as an indicator of image quality. These tests provide reliable data in comparing different SR algorithms because human eyes are the ultimate receivers of the images. However, they are expensive and extremely time-consuming. More importantly, they are difficult to be incorporated into the design and optimization processes of SR and interpolation algorithms.

Automatic or objective image quality assessment (IQA) approaches for image SR is highly desirable but there has been very little progress so far. The difficulty lies in the fact that a perfect-quality HR image is unavailable to compare with. As a result, common IQA approaches such as peak signal-to-noise-ratio (PSNR) and the structural similarity (SSIM) index [5] are not directly applicable.

The purpose of this work is to develop an objective IQA method for a given HR image using the available LR image as a reference. In particular, we take a natural scene statistics (NSS) approach [6], which is based on the hypothesis that the human visual system is highly adapted to the statistics of the natural visual environment and the departure from such statistics characterizes image unnaturalness. In the literature of IQA, such unnaturalness-based measures have been successfully used to evaluate perceived image degradations [6]. In this study, we build statistical models in both spatial and frequency domains and then combine them to produce an overall distortion measure of the HR image. Experimental validation using subjective evaluations demonstrates the effectiveness and usefulness of the proposed algorithm.

2. PROPOSED METHOD

2.1. Frequency Energy Falloff Statistics

It has long been discovered that the amplitude spectrum of natural images falls with the spatial frequency approximately proportional to $1/f^p$ [7], where f is the spatial frequency and p is an image dependent constant. This helps us build a statistical model based on frequency energy falloff. Specifically, we decompose both the HR and LR images into dyadic scales using a steerable pyramid transform [8] (which constitutes a tight frame and thus the energy in the spatial domain is preserved in the transform domain). We then compute the energy (sum of squared transform coefficients) in each scale and observe how the energy falls from coarse to fine scales. An example is shown in Fig. 1, which is computed using an HR "Barbara" image together with an LR version of half size. There are two important observations from this example. First, the falloffs are approximately (but not exactly) straight lines in log-log scale, which is consistent with the $1/f^p$ relationship. Second, the falloffs of the HR and LR images are approximately parallel. These strong structural regularities in the energy falloff curves imply high predictability. In particular, given an LR image, we can compute its frequency energy falloff curve and then use it to predict the *full* falloff curve of its corresponding HR image, even beyond the finest scale in the LR image.

To test the theory motivated from the above discussion, we apply the computation described above to pairs of high-quality LR and HR natural images, and then study how accurately the falloff curves of LR images can predict those of the HR images. We index the scales from coarse to fine so that the finest scales of the HR and LR images

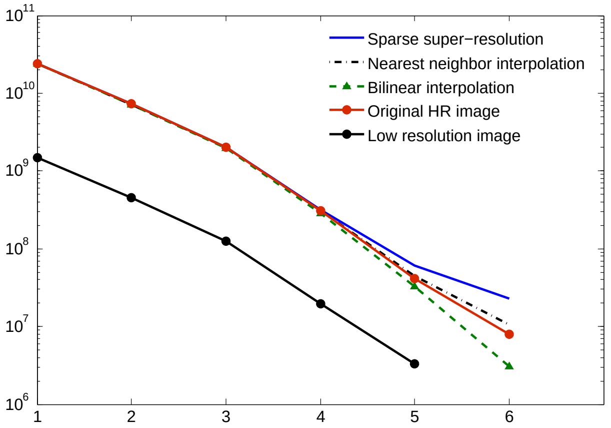

Fig. 1. Frequency energy falloffs of the original LR and HR "Barbara" images as well as interpolated images using bilinear, nearest neighbor, and sparse SR [9] methods.

are Scale 6 and Scale 5, respectively (as exemplified in Fig. 1). Let s_i^H and s_i^L denote the slopes of the falloffs between the i-th and the (i+1)-th scales in the HR and LR images, respectively. To predict s_i^H from s_i^L, we find direct prediction is precise for the first two slopes, i.e., $\hat{s}_1^H = s_1^L$ and $\hat{s}_2^H = s_2^L$. The third and fourth slopes can be well predicted using the following linear models:

$$\hat{s}_3^H = a_0 + a_1 s_3^L \qquad (1)$$

$$\hat{s}_4^H = b_0 + b_1 s_4^L \,, \qquad (2)$$

where the prediction coefficients a_0, a_1, b_0 and b_1 are obtained by a simple least square regression using real high-quality natural images. Once $\hat{s}_3^H$ and $\hat{s}_4^H$ for the HR image are obtained, we can then use them to predict the slope between the finest scales by

$$\hat{s}_5^H = c_0 + c_1 \hat{s}_3^H + c_2 \hat{s}_4^H \,. \qquad (3)$$

Again, the coefficients c_0, c_1 and c_2 here can be obtained using least square regression using high-quality natural images. The prediction coefficients obtained in our regression are given by $a_0 = 0.07$, $a_1 = 1.00$, $b_0 = 0.89$, $b_1 = 1.06$, $c_0 = -3.38$, $c_1 = -0.10$, and $c_2 = 0.89$, respectively. Once all the slopes are predicted, we can then reconstruct a predicted frequency energy falloff curve of the HR image.

When working with the SR quality evaluation problem, the original HR image is unaccessible. The falloffs of the HR images created using SR or interpolation algorithms may be significant different, depending on both the image and the SR/interpolation algorithm. Several examples are shown in Fig. 1, where the largest differences between different methods are observed in the finest scale. This is expected because different SR/interpolation methods have different ways to extend the LR image to finer scales. In particular, the bilinear interpolation method blur the image and thus reduce the high frequency energy, while the nearest neighbors and the sparsity-based SR method [9] add high frequency details to the images, and thus the slopes are raising at the finest scale. Consequently, it is useful to quantify the normalized error in frequency energy falloff between the prediction and the true slope of the HR image at the finest scale:

$$e_f = \frac{\hat{s}_5^H - s_5^H}{\hat{s}_5^H} \,. \qquad (4)$$

Ideally, e_f should be close to zero when the HR image is a high-quality original image. We tested this using 1400 high-quality natural images and the histogram of e_f is shown in Fig. 2, which we find can be well fitted using a generalized Gaussian density (GGD) function

$$p_{e_f}(e_f) = \frac{1}{Z_f} \exp\left[-\left(\frac{|e_f - \mu_f|}{\alpha_f}\right)^{\beta_f}\right] \,, \qquad (5)$$

where $Z_f = \frac{\beta_f}{2\alpha_f \Gamma(1/\beta_f)}$ is a normalization factor, μ_f is the center of the distribution, and α_f and β_f are the width and shape parameters, respectively. This density function becomes peakier at the center with the decrease of β_f. As special cases, $\beta_f = 2$ corresponds to a Gaussian distribution and $\beta_f = 1$ leads to a Laplacian distribution. Our maximum likelihood based fitting result gives $\mu = 0.029$, $\alpha = 0.0608$, and $\beta = 0.6124$, which indicates that the distribution is even peakier than Laplacian. The fitted curve is shown in Fig. 2.

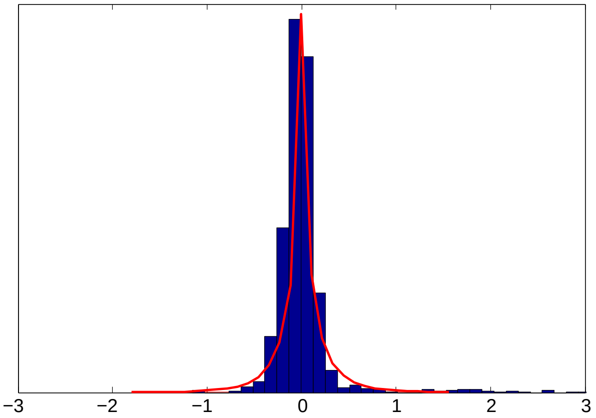

Fig. 2. Histogram and GGD fitting of prediction error e_f of frequency energy falloff for original HR natural images.

2.2. Spatial Continuity Statistics

The above statistical model is fully built in the transform domain. In the spatial domain, interpolation algorithms often create unnatural discontinuities. This motivates us to study continuity based statistical models in the spatial domain and relate them to the naturalness of images. In addition, our method is also inspired by the success of the image blockiness measure proposed in [10].

Let $f(i)$ for $i = 0, \cdots, N-1$ be one row (or column) of pixels extracted from the image, where N is the number of pixels in the row (or column). A straightforward method to examine the signal continuity is to compute an absolute differencing signal

$$g(i) = |f(i+1) - f(i)| \quad \text{for} \quad 0 \le i \le N-2 \,. \qquad (6)$$

In the case of interpolation by a factor of 2, the even and odd samples in $f(i)$ may exhibit different levels of continuities, which will be reflected in the amplitude patterns in $g(i)$. By contrast, such patterns should not be observed in $g(i)$ computed from high-quality natural images. To quantify this, we compute

$$e_s = \frac{1}{M} \sum_{i=0}^{M-1} [g(2i) - g(2i+1)] \,, \qquad (7)$$

where $M = \lfloor N/2 \rfloor$. This spatial continuity measure is computed for every row and every column in the image and then averaged over all rows and columns, resulting in a single overall spatial continuity measure e_s of the whole image. The histogram of the e_s measure of 1400 high-quality natural images is shown in Fig. 3. As in the case of e_f, the histogram can also be well fitted using a GGD model (shown in Fig. 3) given by

$$p_{e_s}(e_s) = \frac{1}{Z_s} \exp\left[-\left(\frac{|e_s - \mu_s|}{\alpha_s} \right)^{\beta_s} \right], \qquad (8)$$

where Z_s is a normalization factor, and the maximum likelihood estimation of the parameters are given by $\mu_s = 0.007$, $\alpha_s = 0.0751$ and $\beta_s = 0.8679$, respectively.

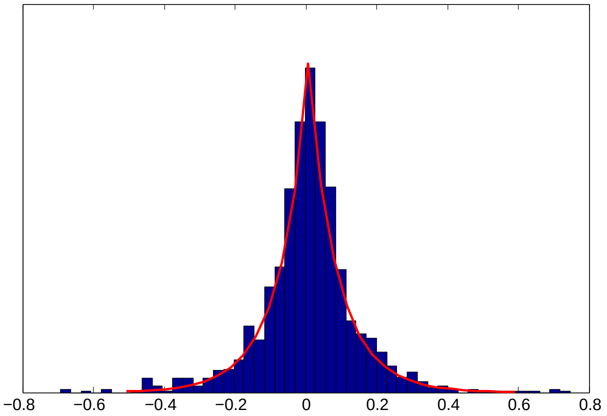

Fig. 3. Histogram and GGD fitting of spatial discontinuity e_s for original HR natural images.

2.3. Quality Assessment Model

The natural image probability models p_{e_f} and p_{e_s} introduced in the previous subsections provide useful measures of the naturalness of images. Based on the statistics we have shown, a high-quality HR natural image should achieve nearly the maximum values in both quantities with high probabilities. An interpolated HR image may depart from such statistics and thus results in lower values. Assume independence of the two probability models, a normalized joint probability measure of naturalness is given by

$$p_n = \frac{1}{K} p_{e_f}(e_f) p_{e_s}(e_s), \qquad (9)$$

where a normalization factor $K = \max\{p_{e_f} p_{e_s}\}$ is added such that the maximum naturalness measure of p_n is up-bounded by 1. It is straightforward to find that

$$K = \frac{1}{Z_f Z_s}. \qquad (10)$$

A commonly used method in information theory to convert this probability-based measure to a "surprisal" based distortion measure is given by

$$D_n = -\log p_n. \qquad (11)$$

Plug (5), (8) and (9) into (11), we have

$$D_n = \left(\frac{|e_f - \mu_f|}{\alpha_f} \right)^{\beta_f} + \left(\frac{|e_s - \mu_s|}{\alpha_s} \right)^{\beta_s} \equiv D_f + D_s, \qquad (12)$$

where we have defined the first term to be the frequency energy falloff feature denoted by D_f and the second the spatial continuity feature by D_s.

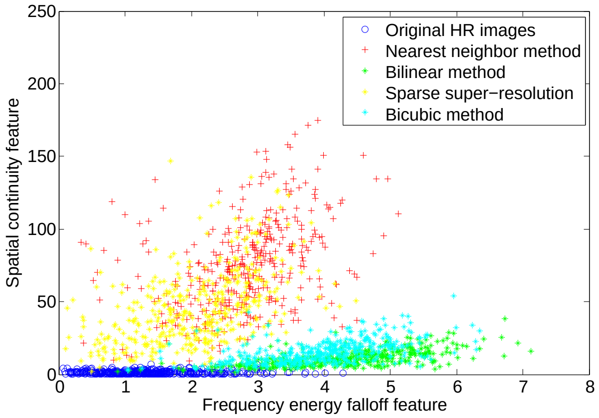

Fig. 4. Scatter plot of natural and interpolated images over frequency energy falloff and spatial continuity features.

Figure 4 shows a scatter plot of 2000 images over the D_f and D_s features, where each point corresponds to one image. These images include both high-quality original HR natural images and HR images created using different SR/interpolation methods. It can be observed that the original natural image cluster is located near the origin. Different SR/interplation methods create different levels of D_f and D_w distortions and are clustered in different locations. For example, the bilinear interpolation method does not have significant spatial discontinuity distortions, but creates severe unnatural frequency energy falloffs (because of its blurring effect). By contrast, the nearest neighbor interpolation algorithm generates blocking artifacts that significantly affect spatial continuity.

Although D_n provides a simple and elegant measure that does not require a training process using any distorted images (all parameters are obtained using high-quality natural images only), it does not take into account the variations in perceptual annoyance to different types of distortions. A natural extension of this approach is to give different weights to difference features. This results in a weighted distortion measure given by

$$D_w = (1 + w) D_f + (1 - w) D_s, \qquad (13)$$

where w determines the relative importance of D_f and D_s, and the special case of $w = 0$ corresponds to the D_n measure. Empirically, we find $w = 0.82$ produces reasonable results in the subjective test discussed in the next section.

3. VALIDATION

A subjective experiment was conducted to validate the proposed algorithm. Twenty subjects were asked to rank 8 image sets, each of which includes 5 HR images generated from the same LR image by 5 different interpolation/SR methods including bilinear, bicubic, nearest neighbor, new edge-directed interpolations [11] and spare representation based super-resolution [9].

To evaluate the proposed measure, we compute the Spearman's rank-order correlation coefficient (SRCC) for each image set between the average subjective rankings and the proposed D_n and D_w

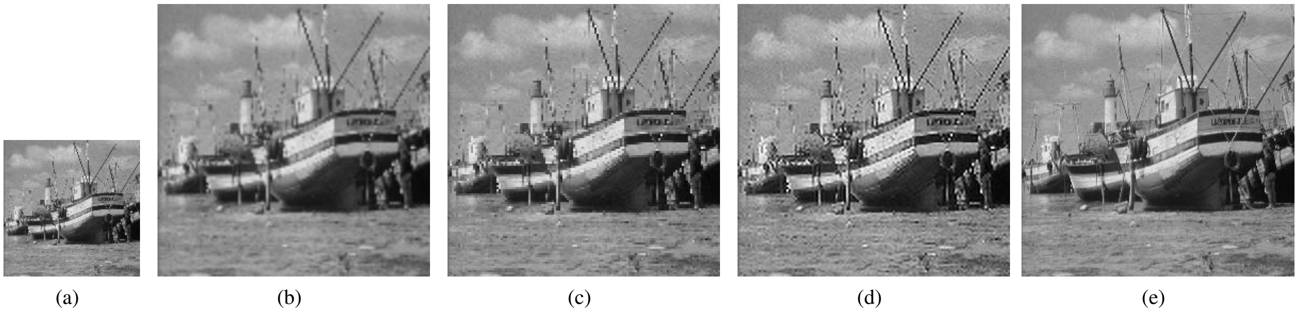

Fig. 5. LR image (a) and the SR/interpolated HR images by (b) bilinear interpolation ($D_f = 3.43$, $D_s = 20.79$, $D_n = 24.22$, $D_w = 9.9$), (c) nearest neighbor interpolation ($D_f = 1.35$, $D_s = 105.5$, $D_n = 106.85$, $D_w = 21.44$), and (d) Sparse SR [9] ($D_f = 2.75$, $D_s = 54.04$, $D_n = 56.79$, $D_w = 14.73$), along with (e) the original HR image ($D_f = 1.01$, $D_s = 0.8$, $D_n = 1.81$, $D_w = 1.9$).

measures. The evaluation results are shown in Table 1. Unfortunately, to the best of our knowledge, no other existing IQA algorithm is applicable to the same scenario and can be included in the comparison. To provide an anchor, we compute the SRCC between the ranks given by each individual subject and the average ranks of all subjects. The mean and standard deviation (std) of SRCC values across all subjects are given in Table 1. This gives an idea about how an average subject behaves in such a test and provides a basis for the comparison of objective methods. In particular, the high std value between subjective opinions reveals that the judgement of the quality of SR/interplation methods is quite difficult even for humans. The proposed D_f and D_s features and the combined D_n measure are positively correlated with the average subjective evaluations while the D_w measure performs significantly better and achieves the same level (or even better) SRCC performance in comparison with an average subject.

Table 1. SRCC evaluation against mean subjective rankings

Average Subject (std)	D_f	D_s	D_n	D_w
0.6515 (0.2868)	0.3125	0.4000	0.4000	0.7125

4. CONCLUSION

We made one of the first attempts to design an NSS-based objective method to assess the quality of HR images created using SR/interpolation methods. Statistical models to capture the naturalness in frequency energy falloff and spatial continuity are constructed and employed in image distortion analysis. Experiments show that the proposed measure agrees well with subjective rankings of overall image quality. The current algorithm is applicable to the case of interpolation/SR by a factor of 2 only. Future work includes extending the current approach for general interpolation factor and investigating other features that could be used to characterize the naturalness of images and to capture the distortions in SR/interpolated images.

5. ACKNOWLEDGMENT

This work was supported in part by Natural Sciences and Engineering Research Council of Canada, Ontario Ministry of Research and Innovation, and Ontario Centre of Excellence, which are gratefully acknowledged.

6. REFERENCES

[1] S. C. Park, M. K. Park, and M. G. Kang;, “Super-resolution image reconstruction: a technical overview,” *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, 2003.

[2] S. Baker and T. Kanade, “Limits on super-resolution and how to break them,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, 2002.

[3] A. R. Reibman, R. M. Bell, and S. Gray, “Quality assessment for super-resolution image enhancement,” in *Proc. IEEE Int. Conf. Image Proc.*, pp. 2017–2020, 2006.

[4] A. R. Reibman and T. Schaper, “Subjective performance evaluation for super-resolution image enhancement,” in *Second Int. Wkshp on Video Proc. and Qual. Metrics (VPQM’06)*, 2006.

[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Proc.*, vol. 13, pp. 35–44, Apr. 2004.

[6] Z. Wang and A. C. Bovik, “Reduced- and no-reference visual quality assessment - the natural scene statistic model approach,” *IEEE Signal Processing Magazine*, vol. 28, pp. 29–40, Nov. 2011.

[7] D. J. Field and N. Brady, “Visual sensitivity, blur and the sources of variability in the amplitude spectra of natural scenes,” *Vision Research*, vol. 37, no. 23, pp. 3367–83, 1997.

[8] E. P. Simoncelli and W. T. Freeman, “Steerable pyramid: a flexible architecture for multi-scale derivative computation,” in *IEEE International Conference on Image Processing*, vol. 3, pp. 444–447, 1995.

[9] J. Yang, J. Wright, T. S. Huang, and Y. Ma, “Image super-resolution via sparse representation,” *IEEE Transactions on Image Processing*, vol. 19, pp. 2861–2873, 2010.

[10] Z. Wang, H. R. Sheikh, and A. C. Bovik, “No reference perceptual quality assessment of JPEG compressed images,” in *IEEE International Conference on Image Processing*, vol. 1, pp. I/477–I/480, 2002.

[11] X. Li and M. T. Orchard, “New edge-directed interpolation,” *IEEE Transactions on Image Processing*, vol. 10, pp. 1521–1527, 2001.