Quality assessment for super-resolution image enhancement

Amy R. Reibman and Sharon Gray
AT&T Labs – Research
{amy,sharon}@research.att.com

Abstract—A typical image formation model for super-resolution introduces blurring, aliasing, and added noise. The enhancement itself may also introduce ringing. In this paper, we assess the quality of images enhanced with super-resolution. We use subjective tests to generate a ground truth. We then examine the ability of some existing quality metrics to quantify the observed subjective quality. Among the metrics tested, SSIM performs best; however, it still does not perform well.

I. INTRODUCTION

Image super-resolution creates an enhanced high-resolution (HR) image using multiple low-resolution (LR) images of the same object. A typical image formation model introduces blurring, aliasing, and added noise. Super-resolution (SR) algorithms jointly reduce or remove all three. The first SR algorithm [1] was based on an image model that only introduced aliasing (no blurring or noise), and explicitly performed de-aliasing in the frequency domain. Subsequently, most SR algorithms have operated in the spatial domain, including iterative back-projection (IBP) [2], projection onto convex sets (POCS) [3], maximum a posteriori (MAP) optimization [4], [5]. Approaches based on wavelets have also been presented [6], [7]. A comprehensive overview of super-resolution algorithms can be found in [8].

One primary drawback of many papers on SR algorithms is that they do not quantify performance. Sometimes, the algorithms are applied to “synthetic” LR images generated from an original HR image. In this case, objective metrics like MSE or PSNR can be computed and are reported. However, when real examples are shown, no HR reference is available. In these cases, performance is typically only evaluated by eye: the SR image is shown and the amount of additional high-frequency detail is described in words.

Recently people have begun exploring the performance of SR algorithms [9], [10], [11], [12], [13] for translational motion. Limits on the upsampling factor for a box-filter PSF have been explored by [9], [10] and [12]. Robinson and Milanfar [11] explore joint performance of aliased image registration and SR image enhancement for integer SR factors. Champagnat and Le Besnerais [13] explore performance using an end-to-end model (formation to reconstruction) in the Fourier domain, also for integer SR factors. However, these all evaluate only objective MSE performance.

In [14], we explore the subjective performance of super-resolution image enhancement. We examine subjective quality as a function of upsampling factor, downsampling factor, and SR algorithm for three images. We show that, in contrast to the claim in [10], image quality improves as the upsampling factor increases. We also show that PSNR does not effectively predict quality of SR images.

In this paper, we extend [14] to examine the ability of some existing quality metrics (including SSIM [15] and Marziliano’s blurring and ringing metrics [16]) to characterize SR performance. We focus on synthetic examples so the original image is available for reference. In addition, we extend our subjective tests to include three more images.

The perceptual impact of image scaling, a problem closely related to super-resolution enhancement, has been explored in [17]. They catalog five impairment types introduced by image scaling. These are unsharpness (or blurring), ringing, and three impairments caused by aliasing. They use expert viewers to judge the severity of each of these five impairment types to obtain an overall quality value for the scaled images. However, the requirement of expert viewers limits the applicability of this method.

In section II, we overview the super-resolution problem. Section III describes our subjective testing methodology, analysis, and our subjective tests. Results of the subjective tests are presented in Section IV. Section V explores six existing quality metrics regarding their ability to quantify quality for this application.

II. SUPER-RESOLUTION IMAGE ENHANCEMENT

We describe our image formation model for super-resolution enhancement using only one dimension for simplicity. The $k$-th observed low-resolution (LR) frame is

$$g_k(n) = [h(x) * f(x - \tau_k) + w_k(x)]_{x=n\Delta}, \quad k = 1, \ldots, K,$$

where $f(x)$ is the continuous image scene, the sampling interval of the LR image is $\Delta$, and $w_k(x)$ is additive white Gaussian noise with variance $\sigma^2$. We assume a translational shift $\tau_k$ between the LR images. The camera applies a continuous point-spread function (PSF) $h(x)$. Super-resolution inverts the image formation process and estimates $f(x)$.

We use the model in [9], [10], [12], where the continuous HR image is histogram-shaped, so the continuous HR image is formed by applying sample-and-hold reconstruction to the discrete samples of the HR image. We also assume that the continuous PSF is space-invariant [12].

Typically, the image formation process is expressed in matrix form. Denote $P_k$ as the matrix operator which combines
downsampling, blurring by PSF $h$, and the shifts $\tau_k$. Then,
\[ g_k = P_k f + w_k, \quad k = 1, \ldots, K \tag{2} \]

where $g_k$ are the LR images, $f$ is the high-resolution image, and $w_k$ are the noise vectors. Because we are interested in non-integer upsampling factors, the sampling grid of the LR image is typically not a subset of the sampling grid on which we want to estimate $f$. Therefore, rows of $P_k$ are not shifted versions of each other. Specifically, we use the process described in [12] to compute a discrete PSF – on the HR grid to be applied to the HR image – for each pixel in the LR images, dependent on its offset relative to the HR grid.

$L$ is the upsampling factor; the factor by which the SR image $\hat{f}$ is larger than the LR images $g_k$. $M$ is the synthetic downsampling factor; the factor by which we downsample the original image (after filtering) to obtain the LR images $g_k$. Algorithmic details for MAP, the algorithm we use, can be found in [8] and the references therein.

III. SUBJECTIVE TEST METHODOLOGIES

A. Experimental test design

In [14], we examine three monochrome images: Fingerprint, Tiger, and the text image ReadThis. In this paper, we add three new monochrome images: Arch, Girl, and Railing. These images all make aliasing and blurring readily apparent. The original image Arch is shown in Figure 1(a).

We displayed all images on a Nokia 445XiPlus CRT monitor at a viewing distance of 24 inches. The task for each viewer was to select which image from a pair of images was “better”. Each experiment in this paper uses twenty-six viewers.

In each set of images, the pairs to test were chosen adaptively, in a manner similar to that proposed by Silverstein and Farrell [18] which ensures that more similar pairs are compared more often. We also add a final step for viewers to compare all pairs that are separated by one in the sorted list, which adds more tests of similar images without significant additional tests [14].

B. Test 1: Downsampling by six, variable upsampling

The goal of the first experiment is to determine if there is a visual limit of super-resolution. Also, the secondary goal is to determine if viewers rank images using integer SR factors as better/worse or the same as images using a nearby (lower) non-integer SR factor.

Each original image is processed as follows. First, we apply a $6 \times 6$ uniform blur (i.e., a box-filter of size 6 pixels). Second, we apply different offsets and downsampling by $M = 6$, to obtain $K = 36$ low-resolution (LR) images. Because the $M = 6$ box-filter has even-length support, the offsets of the LR images are in the set $\{1/(2M), \ldots, 1-1/(2M)\}$. No noise was added. An example LR image, interpolated to the same size as the original, is shown in Figure 1(b).

We apply MAP super-resolution to the 36 LR images examining 10 different upsampling SR factors $L \in \{2, 3, 4, 5, 6, 7, 8\}$ and $L = M$.

Figure 1 shows the Arch image (a) original, (b) one LR image bilinearly interpolated, (c) SR with $M = 6$ and $L = 2$ (box-PSF) (d) SR with $M = 6$ and $L = 4$ (box-PSF). These images are illustrative only; rescaling during printing alters the perception of the images.

We initialize the algorithm using bilinear interpolation. We iterate 500 times with $\alpha = 0.01$. We use ML optimization for all SR factors; no regularization is applied, even for integer SR factors. After the ML-SR algorithm, we use cubic interpolation to scale the output from each SR factor to an overall scale factor of $L = 6$. Thus, each pair compared in the subjective test has the same size.

Figure 1 shows the Arch image (a) original, (b) one LR image with bilinear-interpolation to original size (c) the SR result with $L = 2$ and (d) the SR result with $L = 4$. Aliasing artifacts are evident when $L = 2$, and ringing artifacts are present both when $L = 2$ and $L = 4$.

C. Test 2: Variable downsampling, upsampling to original size

In our second test, we generate synthetic LR images using different degrees of downsampling, while the upsampling factor in each case returns the image to the spatial resolution of the original. The PSF is again the box-filter, with size $M$ according to the downsampling factor. Here, $M \in \{2, 3, 4, 5, 6, 7, 8\}$ and $L = M$.

In addition, to obtain relative quality between the two sets of images, we compare the four best images from Test 1 to the four worst images from Test 2. All viewers compared each of the sixteen combinations.
D. Statistical analysis

At the suggestion of Handley [19] we use the Bradley-Terry model [20] to analyze the subjective test results. The goal is to find the quality, $Q_i$, for each of the $t$ images under test. In the Bradley-Terry model, the relative quality between two images is $Q_i - Q_j = \log \pi_i - \log \pi_j$, where the probability that image $i$ is preferred to image $j$ is modeled by $\pi_i / (\pi_i + \pi_j)$, where $\sum_1^t \pi_i = 1$. This model only defines the relative quality between tested images, not absolute quality.

The maximum of the likelihood function corresponds to the quality values that would produce the observed outcome (i.e., the subjective test results) with the highest probability. If we let $n_{ij}$ be the number of times pair $(i, j)$ is compared and $a_i$ be the total number of times image $i$ was preferred relative to all images, then the maximum likelihood estimate, $p_i$, of $\pi_i$ is

$$p_i = a_i / \left( \sum_{j \neq i} n_{ij} / (p_i + p_j) \right), \quad i = 1, \ldots, t. \quad (3)$$

These $t$ equations can be solved iteratively [20]. The Bradley-Terry model also offers confidence intervals on the quality values and hypothesis tests. Details can be found in both [20] and [19].

IV. SUBJECTIVE TEST RESULTS

A. Test 1: Downsampling by six, variable upsampling

Figure 2 shows the maximum likelihood (ML) estimates for the first test if we combine the results from all six images. Quite clearly, subjective quality continues to improve as the SR factor $L$ increases. However, with all images, we cannot reject the hypothesis “integer SR factors are not better than nearby lower non-integer SR factors” with 95% confidence. If we exclude the Arch image, however, we can reject this hypothesis for the comparison between $L = 1.9$ and $L = 2$.

Figures 3 show the results from individual images. The subjective quality on the y-axis is relative for each curve and should not be compared across curves; in fact, we add an offset between the qualities to make the graph easier to read. Viewers had a hard time distinguishing among the different images for Girl; however for all other images, quality generally improves as the upsampling factor $L$ increases. The increase is not always monotonic; some dips that lend support to the claim in [10] are for Arch ($L = 2$) and Fingerprint ($L = 3$). Peaks that contradict the claim are for Girl and Tiger ($L = 2$).

Due to space constraints, we leave analysis of Test 2 to the next section, where we examine the behavior of the quality metrics.

V. QUALITY METRICS

In this section, we examine the ability of 6 metrics to quantify super-resolution quality. These are MSE, Weighted SNR as described in [21], SSIM [15], Marziliano's blurring and ringing metrics [16], and JP2kNR, a no-reference metric for JPEG-2000 [22]. Among these, only the blurring metric and JP2kNR are no-reference metrics. While JP2kNR is designed for a different application, we explore its applicability here since both SR and JPEG-2000 have both blurring and ringing artifacts.

Because our ground truth is only relative within each image, we only examine how well these metrics can predict the rank order of quality. This ability is critical if we compare different processing results on the same original image. We use Kendall’s tau $\tau_a$ [23], which is an estimate of the probability that a pair of variables is more likely to be correctly ordered than incorrectly ordered. For completely ordered data, $\tau_a = 1$. As pairs become incorrectly ordered, $\tau_a$ decreases.

We perform a joint analysis across the images in both Test 1 and Test 2. For each original image, we compute one relative quality scale. Kendall’s tau is given in Table I for each of five images and the six metrics.

All metrics fail on Girl, since the viewers themselves had a hard time distinguishing among images. MSE has better than 90% accuracy on two of the five images, and better than

1In [14] we did not test ReadThis in Test 2.
80% accuracy on two others. The blurring metric only beats MSE for Arch, while the ringing metric only outperforms MSE for Girl. The best performance is obtained by SSIM, which performs the same as MSE for three images but edges out MSE for the other two. Even so, it obtains greater than 90% accuracy on only two of the five images.

We show the relative quality as a function of the SSIM for five images in Figure 4, to highlight the difficulties of accurate quality assessment for SR. (The relative quality should not be compared across images.) The solid lines connect the images in each test consecutively, and the dotted lines connect the case \( M = 6, L = 4 \) with \( M = L = 6 \). Test 1 with \( L = 1.9 \) has the lowest SSIM for every image, and Test 2 with \( L = M = 2 \) has the highest SSIM for every image.

Behavior for Fingerprint is nearly ideal; SSIM increases with subjective quality smoothly, even between the two tests. However, SSIM cannot predict the decrease in quality for Test 1, \( L = 3 \). For Railings and Arch, SSIM cannot predict the non-monotonic subjective quality within both Test 1 and Test 2, but the results across the tests blends smoothly.

However, for Tiger and Girl, SSIM does not perform as well between the two tests. There is little subjective improvement for Tiger between \( M = 6, L = 4 \) in Test 1 and \( M = L = 6 \) in Test 2; however, SSIM has a sizeable increase. In addition, for both Tiger and Girl, SSIM does not correctly predict the quality of \( L = M = 7 \) and 8 for Test 2 relative to the Test 1 images. We would expect a single smooth trajectory independent of the test, which does not happen.

These results demonstrate that a deeper understanding of the subjective performance of super-resolution is essential. The image formation process introduces blurring, aliasing, and noise, and super-resolution enhancement may also introduce ringing. However, existing ringing and blurring metrics are not adequate to characterize subjective performance, since they cannot account for aliasing.

### TABLE I

<table>
<thead>
<tr>
<th>Image</th>
<th>MSE</th>
<th>WSNR</th>
<th>SSIM</th>
<th>Blur</th>
<th>Ring</th>
<th>JP2kNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arch</td>
<td>0.83</td>
<td>0.83</td>
<td>0.87</td>
<td>0.88</td>
<td>0.77</td>
<td>0.98</td>
</tr>
<tr>
<td>Girl</td>
<td>0.45</td>
<td>0.58</td>
<td>0.45</td>
<td>0.35</td>
<td>0.52</td>
<td>0.30</td>
</tr>
<tr>
<td>Railings</td>
<td>0.93</td>
<td>0.90</td>
<td>0.93</td>
<td>0.78</td>
<td>0.87</td>
<td>0.85</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>0.98</td>
<td>0.90</td>
<td>0.98</td>
<td>0.35</td>
<td>0.13</td>
<td>0.35</td>
</tr>
<tr>
<td>Tiger</td>
<td>0.82</td>
<td>0.90</td>
<td>0.85</td>
<td>0.80</td>
<td>0.52</td>
<td>0.70</td>
</tr>
</tbody>
</table>

![Figure 4](image_url)  

Fig. 4. ML estimates of image quality, Tests 1 and 2, as a function of SSIM. Relative quality should be compared only within a curve, not between curves.

### REFERENCES