A No-Reference Metric for Evaluating the Quality of Motion Deblurring

Yiming Liu[1]* Jue Wang[2] Sunghyun Cho[2] Adam Finkelstein[1] Szymon Rusinkiewicz[1]
[1]Princeton University [2]Adobe Research

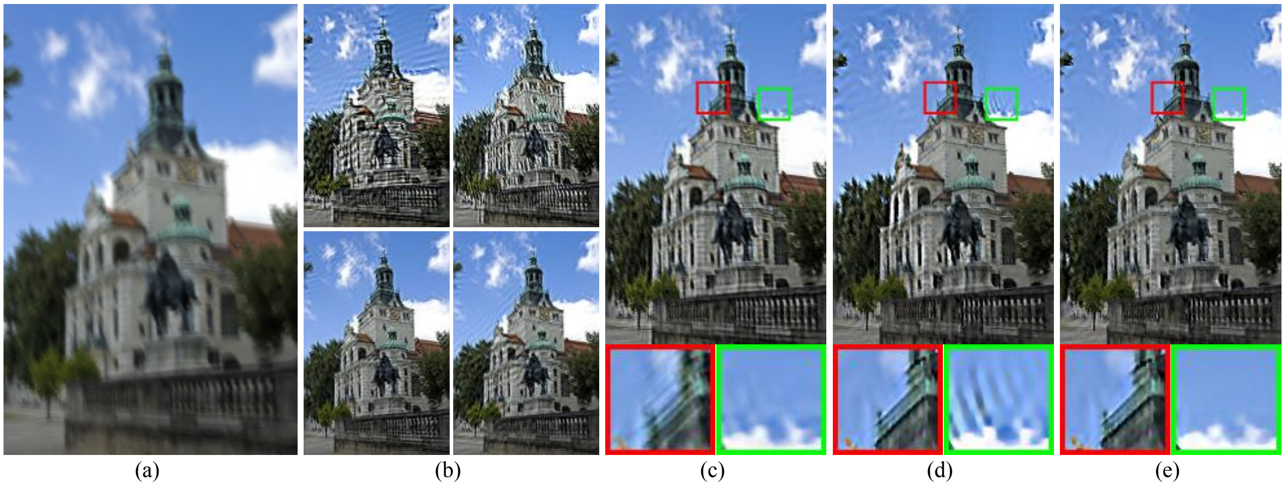

Figure 1: *We develop a no-reference metric for evaluating the perceptual quality of image motion deblurring results. The metric can be used for fusing multiple deblurring results (b) of the same input image (a) to generate one with the best quality (e). (c) and (d) are the result of simply averaging all deblurring results and the result of a naive fusion method, respectively. See text in Sec. 6.3 for more details. Original image courtesy digital cat@Flickr.*

Abstract

Methods to undo the effects of motion blur are the subject of intense research, but evaluating and tuning these algorithms has traditionally required either user input or the availability of ground-truth images. We instead develop a metric for automatically predicting the perceptual quality of images produced by state-of-the-art deblurring algorithms. The metric is learned based on a massive user study, incorporates features that capture common deblurring artifacts, and does not require access to the original images (i.e., is "no-reference"). We show that it better matches user-supplied rankings than previous approaches to measuring quality, and that in most cases it outperforms conventional full-reference image-similarity measures. We demonstrate applications of this metric to automatic selection of optimal algorithms and parameters, and to generation of fused images that combine multiple deblurring results.

CR Categories: I.3.0 [Computer Graphics]: General—;

Keywords: image quality metric, no-reference, percetually-validated, deblurring

Links: DL PDF Web

1 Introduction

The wide availability and ever-increasing sophistication of modern image processing and computational photography algorithms has brought about a need to evaluate their results. For instance, for a task such as image deblurring, a realistic characterization of image quality and the presence or absence of artifacts is necessary to select between different methods, as well as to choose parameters for each algorithm. Lacking an automated method for image quality assessment, many systems resort to asking the user. This, however, becomes increasingly impractical if dozens of algorithms and hundreds of parameter settings must be compared. While a large-scale user study (using, for example, the Amazon Mechanical Turk) might be able to compare many combinations of algorithms and parameters, it would be unrealistic to use this methodology for *every* image that is processed. Finally, the lack of a high-quality "ground truth" image in most applications precludes the use of traditional metrics for image comparison, such as peak signal to noise ratio (PSNR), structural similarity index (SSIM), or visual information fidelity (VIF).

We explore a methodology in which image quality and artifacts are scored according to some function of features that are computed over the image. The function is learned based on thousands of training examples provided in a massive online user study. By picking the right features, and collecting enough user input, we are able to obtain a metric that generalizes over images and over the algorithms used to process them. We call this learned function a *perceptually-validated metric*, as it is not built upon the underlying psycho-physical mechanisms of the human visual system, but rather can score image quality and artifacts consistently with human ratings.

In this paper, we address the problem of motion deblurring or blind deconvolution: undoing the (unknown) motion blur introduced by camera shake. This is a problem that has been studied intensively over the past several years [Fergus et al. 2006; Yuan et al. 2007; Shan et al. 2008; Cho and Lee 2009; Krishnan and Fergus 2009;

*Part of this work was done when the first author was an intern at Adobe Research.

Whyte et al. 2010; Xu and Jia 2010; Cho et al. 2011; Levin et al. 2011; Zoran and Weiss 2011; Goldstein and Fattal 2012], but even state-of-the-art methods may produce significant artifacts such as noise and ringing. In surprisingly many cases, these algorithms may produce even worse results than the input blurry image. This makes it crucial to identify which methods are performing well on which images.

The main contribution of this paper is a perceptually-validated metric for evaluating the output of deblurring algorithms. Our key hypothesis is that an image quality metric specialized to this problem will outperform general metrics, or ones specialized to different problems. We therefore conduct a crowd-sourced user study (Sec. 3), incorporating five deblurring algorithms applied to hundreds of images, to learn the relative importance of the principal artifacts of blind image deconvolution: ringing, noise, and residual blur. We design features (Sec. 4) to measure these artifacts, and use them to obtain a mapping from a feature vector to image quality (Sec. 5) that matches the users' rankings as closely as possible. We find that in most cases this *no-reference* metric (i.e., not having access to the ground-truth image, which is important for real-world applications) outperforms existing no-reference image quality metrics, as well as standard *full-reference* image comparison algorithms.

Our user study yields interesting conclusions about the relative importance of different artifacts. For example, we find that large-scale "ringing" is overwhelmingly harmful to perceived image quality, to a much greater extent than noise and blur. These kinds of conclusions may have influence on the design of future algorithms for deconvolution, or even other image processing tasks. The resulting metric also enables applications (Sec. 6) including automatic selection of the best deblurring algorithm for a given image, automatic parameter selection, and fusion of the highest-quality regions of different deblurring results.

We summarize our contributions as:

- We learn a perceptually-validated metric for measuring the quality of image deblurring.
- We provide to the community a data set with specially designed input images, current state-of-the-art deblurring algorithms' results, and users' feedback about their quality.
- We show applications that utilize our metric to automatically produce an improved deblurring result.
- We analyze the impact of each artifact on perceptual quality, which can guide the direction of future work and development of deblurring algorithms.

2 Background and Related Work

Image deblurring. Motion blur caused by camera shake is a common problem observed in photos captured by hand-held cameras. A blurred image b may be modeled as

$$b = l * k + n, \tag{1}$$

where l is a sharp latent image, k is a point spread function (PSF), or a blur kernel, reflecting the trajectory of the camera shake, and n is noise. Deblurring is the problem of solving for a sharp latent image l, given a blurred image b. If k is known, the problem is called *non-blind deconvolution*; otherwise it is called *blind deconvolution*.

Even non-blind deconvolution is an ill-posed problem, since the PSF usually contains null frequencies and the noise level is unknown. To resolve the ambiguity, different types of prior knowledge are used. Levin et al. [2007] propose a natural image prior, which approximates the heavy-tailed distribution of image gradients. Joshi et al. [2009] assume that within each small patch, colors should be linear combinations of two colors. Zoran and Weiss [2011] exploit patch priors trained on natural images.

Blind deconvolution (with unknown PSF) is even more ill-posed, requiring applying strong priors on both the latent image and the PSF. Fergus et al. [2006] assume that the gradients of natural images follow a heavy-tailed distribution and the PSF has a sparse support, and adopt a variational Bayesian approach to estimate a PSF. Levin et al. [2011] assume similar properties on the gradients of natural images, but use an expectation-maximization (EM) method for optimization. Several recent approaches [Joshi et al. 2008; Cho and Lee 2009; Xu and Jia 2010; Cho et al. 2011] rely on extracting edge profiles for PSF estimation, under the assumption that edges are sharp in natural images. Shan et al. [2008] exploit the sparsity on both the latent image and the PSF, and also a local smoothness prior to reduce ringing artifacts. Goldstein and Fattal [2012] exploit the power-law dependence of the power spectra of natural images for PSF estimation.

Despite the tremendous progress that has been made in recent years, state-of-the-art deblurring methods still tend to generate noticeable visual artifacts on real-world data, such as ringing, noise, residual blur, etc. These artifacts may arise because: (1) the priors used in existing methods are simplified approximations to natural image statistics, thus may lead to estimation errors in the PSF and the recovered image; (2) non-linear response curves and non-additive noise in real-world images violate the linear blur model in Eqn. 1; and (3) real-world images often contain spatially-varying blur that may not be well approximated by a static PSF. There has been work on estimating spatially-varying PSFs [Whyte et al. 2010; Gupta et al. 2010; Ji and Wang 2012]; however, as the dimensionality of the PSF increases, these methods are often less reliable and may produce more severe artifacts, as shown in a recent study [Köhler et al. 2012]. Schuler et al. [2012] detect spatially-varying PSFs for images with optical aberrations. However, their method strongly depends on the symmetry properties of optical aberrations, thus is not general enough for handling motion blur.

Crowd-source analysis. In recent years, researchers have analyzed problems related to perception using the data collected from large-scale user studies. For example, Cole et al. [2009] study human perception of line drawings depicting 3D shapes based on a gauge-figure study. Chen et al. [2009] collect 3D segmentation results from a user study, and analyze evaluation criteria based on them. Secord et al. [2011] conduct a large study of viewpoint preferences and develop a model for evaluating views of 3D models. These approaches adopt the strategy of learning perceptual models or criteria from a large-scale user study, then using them for producing or evaluating visual results in a way that tries to match human judgments. In this work we apply this strategy to a new problem: motion deblurring.

Image quality metrics. Image quality metrics can be classified into two categories: full-reference and no-reference, depending on whether the ground truth image is needed. Commonly used full-reference metrics include PSNR [Teo and Heeger 1994], multi-scale SSIM [Wang et al. 2004], VIF [Sheikh and Bovik 2006], HDR-VDP-2 [Mantiuk et al. 2011], and a linear combination of them [Masia et al. 2012]. Although these metrics have been widely used for evaluating subtle image corruption, they are not able to measure the perceptual impact of the significant artifacts introduced by a variety of graphics and vision algorithms, such as image deblurring, as we will demonstrate later, and photo-realistic rendering, as demonstrated by Cadik et al. [2012].

On the other hand, many no-reference metrics, such as BIQI [Moor-

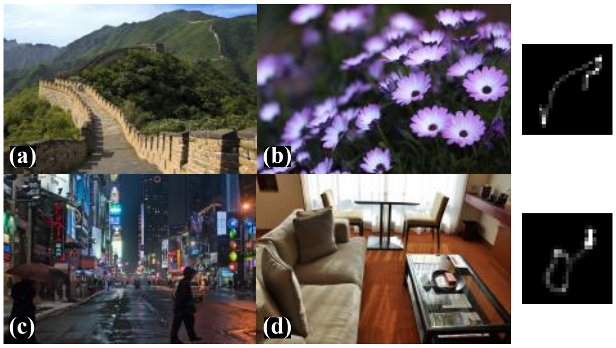

Figure 2: *Left: four ground truth images used in our user study, which contains 40 images in total. Right: two PSFs used in the study. Image courtesy (a) Craig Maccubbin, (b) Jun Seita, (c) Bob Jagendorf, (d) uggboy@Flickr.*

thy and Bovik 2010], BLIINDS [Saad et al. 2010], CORNIA [Ye et al. 2012], LBIQ [Tang et al. 2011], and BRISQUE [Mittal et al. 2012c], use either supervised or unsupervised learning on a collection of images and their subjective scores to produce a general image quality metric. NIQE [Mittal et al. 2012b] uses unsupervised learning without subjective scores. However, unlike our metric, these metrics are not designed for the specific problem of image deblurring, which generates unique artifacts such as strong ringing.

Image quality metrics have been successfully applied in a variety of applications, such as automatic parameter selection [Zhu and Milanfar 2010; Mittal et al. 2012a], blur-aware and noise-aware downsizing [Samadani et al. 2010; Trentacoste et al. 2011], and coded aperture design [Masia et al. 2012]. We will discuss the difference between these methods and our approach in more detail in Sec. 4.2 and 6.1.

3 Data Sets and User Studies

To understand how human beings perceive different deblurring artifacts, we first conduct a massive crowd-sourced user study to evaluate the perceptual quality of deblurring results. Here we describe the data set and the user study in more detail, then discuss some major observations we draw from the user study results, which serve as guidelines for our feature design in Sec. 4.

3.1 The Data Set

Our data set consists of synthetically motion-blurred images and the deblurring results generated by different algorithms. The deblurring results contain a large variation in quality: we intend to include good results as well as ones that have significant artifacts of various kinds.

Specifically, we begin with 40 sharp, high quality images of various scenes as the ground truth, which cover a wide variety of common scenes, such as landscape, cityscape, portrait, and indoor scenes. As several deblurring algorithms rely on extracting edges in the images, we also include images with different amounts of structural edges to have various levels of deblurring difficulty. Fig. 2 shows four of them. We downsample them so that their longest edges are 768 pixels. We then synthetically blur them with two different PSFs shown in Fig. 2, whose sizes are 27×27 and 23×23. In practice, the first PSF is more difficult to deal with and is more likely to cause ringing artifacts. The second PSF usually leads to relatively good results with more subtle artifacts. They lead existing methods to produce results with different levels of artifacts. Finally, we add Gaussian noise of three different levels ($\sigma = 0.0, 0.01, 0.02$) to each blurred image, resulting in $40 \times 2 \times 3 = 240$ blurred examples in total.

We run five recent algorithms [Fergus et al. 2006; Shan et al. 2008; Cho and Lee 2009; Levin et al. 2011; Goldstein and Fattal 2012] with publicly-available executables on all the blurred images to generate deblurring results. In doing so, we encountered some program failures (roughly 0.08% of all test cases) due to the instability of the research prototypes. For each sharp image, we obtain at most 30 deblurring results in this way. We call the collection of each sharp image and its deblurring results a *data group*.

3.2 The User Study

We employ the Amazon Mechanical Turk (MTurk) for the crowd-sourced user study, in which we ask users to compare and rank the quality of deblurring results.

There are a few options for constructing such a study. The most straightforward approach would be to ask each user to give a score for each deblurring result. However, absolute scores are subjective and inconsistent across users. Another choice would be to show all results at the same time and ask the user to rank all of them. This would solve the inconsistency issue, but the large number of images in each data group would make this task tedious for users.

To avoid these problems, in our study we ask users to compare the quality of deblurring results *pairwise*. For each comparison a pair of images is shown side-by-side, and the user is requested to choose the one that has better visual quality (thus using a forced-choice methodology). Each user session consists of 40 pairs of images in total. Once all the pairwise comparison results are obtained, we use them to fit a global ranking, as described in detail in Sec. 3.3.

To ensure the quality of the user study results, our user study includes a mechanism to detect and reject "bad" results produced by malicious or careless users. For each data group, we include the ground-truth sharp image in the user study. Specifically, among the 40 comparisons a user performs in one session, 13 of them include ground-truth images, which are considerably better than most deblurring results. We reject results from users who ranked the ground truth images lower than the deblurring results more than twice. Under this design, the probability that randomly selecting an image in each pair will pass the test is 1.12%, which means that we can effectively reject outliers in the data.

In our study we collected $13,592$ user sessions from $1,041$ users, and 4% of them were rejected by the sanity check. In the remaining data, each pair was ranked by at least 20 different users.

3.3 Fitting a Global Ranking

The problem of fitting pairwise comparison results to a global ranking has been well studied. We adopt the Bradley-Terry model [Bradley and Terry 1952], which is widely used for this purpose. The Bradley-Terry model generates a global "score" for each data point from the pairwise comparison. Here we briefly review how this model works. For each pair of images A and B, assuming their scores are δ_A and δ_B, let $\delta_{AB} = \delta_A - \delta_B$. We can then define the relation between the probability p_{AB} that one user chooses A over B, and the score difference δ_{AB} as:

$$p_{AB} = \frac{e^{\delta_{AB}}}{1 + e^{\delta_{AB}}} = \text{logit}^{-1}(\delta_{AB}), \qquad (2)$$

where $\operatorname{logit}(p) = \log\big(p/(1-p)\big)$. Here we see that:

$$p_{AB}+p_{BA} = \frac{e^{\delta_{AB}}}{1+e^{\delta_{AB}}}+\frac{e^{\delta_{BA}}}{1+e^{\delta_{BA}}} = \frac{e^{\delta_{AB}}}{1+e^{\delta_{AB}}}+\frac{1}{1+e^{\delta_{AB}}} = 1. \tag{3}$$

In the user study, since multiple users have compared A and B, we denote the number of users who favor A as a, and the number of users who favor B as b. Assuming the decisions are independent, a should follow a binomial distribution $\operatorname{Bin}(a+b, a, p_{AB})$. Therefore, the likelihood of (δ_A, δ_B) is:

$$L(\delta_A, \delta_B) = P(a, b \,|\, \delta_{AB}) = \binom{a+b}{a} {p_{AB}}^{a}\,(1-p_{AB})^{b}. \tag{4}$$

The likelihood for all pairs (A, B) is:

$$L \;=\; \prod_{(A,B)} \binom{a+b}{a} {p_{AB}}^{a}\,(1-p_{AB})^{b}, \tag{5}$$

which is a function of δ_{AB}. Note that L only encodes the *difference* between scores of images. In other words, L does not change if the same offset Δ is added to all δ. To resolve this ambiguity, we need a reference image R with $\delta_R = 0$. In our user study, we let all ground truth images be the reference images. We then solve for δ of all images by maximizing L.

3.4 Observations

Based on the user study results, we make the following qualitative observations that provide useful insights for developing our metric:

- The main artifacts that appear in deblurring results include noise, ringing, and blurriness, and these artifacts typically coexist in a deblurred image.
- In general users are very sensitive to ringing artifacts, which unlike noise and blurriness do not exist in natural images. The lowest-ranked images often contain strong ringing artifacts.
- The characteristics of the noise introduced by different deblurring algorithms can be very different.

Fig. 3 shows an example of deblurring results from one data group.

We use these observations as general guidelines for designing low-level image features used to train the perceptually-validated metric, as detailed in the next section.

4 The Collection of Features

We now derive a set of low-level image features for assessing the quality of deblurred images. These features serve as the basis for learning our metric. In general, our expectations for good deblurring results are twofold: (1) *naturalness*: the deblurred image should have a natural appearance; and (2) *sharpness*: the deblurred image should be sharp and have as little residual blur as possible. We design a collection of features to measure how well a deblurring result meets each constraint. Note that not all features will be eventually used in the perceptually-validated metric, and we will discuss how to select good features in Sec. 5.4.

4.1 Measuring Image Naturalness

The naturalness of a deblurring result describes the extent to which it looks like a natural image captured by a camera. While it is hard to measure naturalness directly, it is easier to describe artifacts that often appear in deblurring results making them look unnatural. We identify two dominant artifacts, which are common in deblurring results — *noise* and *ringing* — and design our feature set with a various methods to measure them quantitatively.

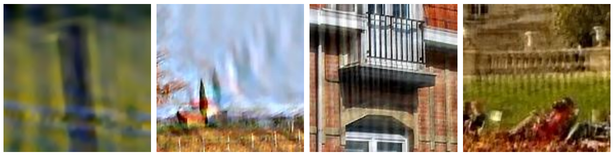

Figure 4: *Typical ringing artifacts in deblurred images.*

4.1.1 Noise

There exists a vast amount of previous work on estimating and removing image noise. Nevertheless, accurate noise estimation remains a challenging problem. Given that different noise estimation methods may work for different types of noise (e.g. chromatic noise vs. luminance noise), we use multiple measures to assess the amount of noise in a deblurred image. We briefly describe each measure below. We refer the readers to the supplementary material for details.

- **Two-color priors** [Joshi et al. 2009] assume that, in natural images, colors within a local image neighborhood are linear combinations of a primary and secondary color. Following [Joshi et al. 2009], we find the two prevalent colors for each local neighborhood, and measure the noise of each pixel based on the two-color model.
- **Sparsity priors (Sps)** [Levin et al. 2007] assume that gradient magnitudes of natural images follow a heavy-tailed distribution. We measure the norm of gradient magnitudes.
- **Gradients of small magnitudes (SmallGrad)** usually correspond to noise on flat regions. We use the variance of the top $m\%$ of smallest gradient magnitudes as a noise measure.
- **MetricQ** [Zhu and Milanfar 2009] is a no-reference metric sensitive to both noise and blur, which is based on the fact that noise and blur make an anisotropic patch more isotropic.
- **BM3D** [Dabov et al. 2007] is one of the state-of-the-art denoising methods. It takes a parameter σ that specifies the noise level in the input image. We estimate the optimal value of σ and include it in our noise feature set. Specifically, we apply BM3D to the image with different values of σ, and measure the errors of the results using a two-color prior. We choose the smallest σ, whose error is smaller than a certain threshold, as a noise measure.

4.1.2 Ringing

Ringing is perhaps the most common artifact one may observe in a deblurred image. It often appears as un-natural wavy structures parallel to sharp edges (Fig. 4). Strong ringing artifacts may span large image regions, or even the entire image. They are mainly caused by inaccurate PSF estimation, but even with an accurate PSF, they can still appear due to the Gibbs phenomenon [Yuan et al. 2007].

Since we have observed that ringing is one of the most annoying artifacts in deblurring (Sec. 3.4), detecting and quantifying ringing is crucial for assessing the quality of a deblurred image. Unfortunately, detecting ringing artifacts is not easy, as ringing is a mid-frequency signal, which is mixed together with the image signal in the same frequency band. It is especially hard to detect ringing in textured regions where rich mid-frequency image signal

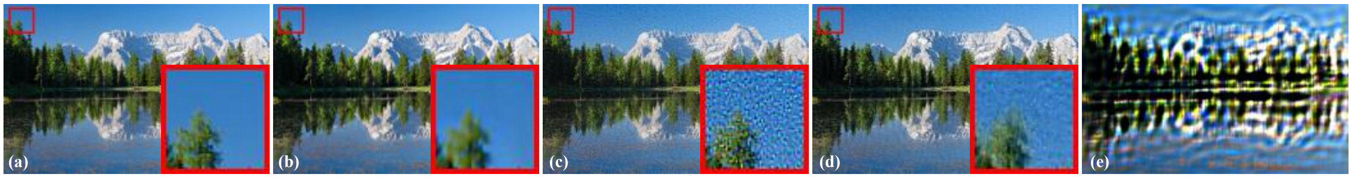

__Figure 3:__ Sample deblurring results from one data group. (a) ground truth; (b) a deblurring result with blur; (c, d) two deblurring results with the same noisy input but different deblurring algorithms, yielding different types of noise in the results; (e) the lowest-ranked deblurring result, which contains strong ringing. Please refer to the supplementary materials for the ranking of the full data group. Original image courtesy Umberto Salvagnin.

exists. Methods have been proposed for measuring ringing artifacts [Marziliano et al. 2004], but most of them are only applicable to small scale ringing caused by lossy compression such as JPEG compression, and/or are full-reference metrics.

We introduce a new method to measure large-scale ringing caused by motion deblurring. We design our method to be conservative: it can reliably estimate ringing in flat regions, but is more conservative in textured regions to avoid misclassifying image structures as ringing, which will significantly damage the image quality assessment. It is also consistent with human perception, as ringing artifacts are more noticeable in smooth regions (Fig. 4).

Full-reference ringing detection. We first describe a full-reference method, which forms the basis for the no-reference method described later. The method begins with a deblurred image l' and the ground truth l, and applies the following steps:

1. We first align l and l' and compute the gradient maps of l and l' as $g(l)$ and $g(l')$.
2. We compute the difference map $\delta g = \max(g(l') - g(l) * k_g, 0)$. k_g is a scaled Gaussian function with a maximum value of 1, which is convolved with $g(l)$ to avoid misclassifying residual blur as ringing artifacts. We take $\max(\cdot)$ because most ringing artifacts in l' should have larger gradients than l;
3. Then, we compute the average of δg as a measure of ringing artifacts in l'.

No-reference ringing detection (PyrRing). For no-reference detection, we substitute l with the input blurry image b. As we apply a low-pass filter in step 1, the error caused by this substitution is not significant. To further mitigate the error, we create an image pyramid for both b and l', and measure the ringing artifacts on each level of the pyramid, then compute the average across all levels as the final indicator of ringing artifacts. At coarse levels, the downsized blurry input b will be closer to the downsized ground truth l, leading to more accurate ringing detection. Thus, the final ringing measurement computed from the pyramid is also more accurate than the one directly computed from the original l' and b. Fig. 5 shows an example of our ringing detection.

Saturation. The ringing artifacts in images are usually accompanied with overshoot and undershoot artifacts. We measure the proportion of pixels with pixel value below 10 or above 245 (assuming all pixel values are between 0 and 255) as the saturation feature.

4.2 Sharpness

An ideal deblurred image should be sharp and contain no residual blur. On the other hand, an inaccurate PSF or too-strong regularization in deconvolution may cause remaining blur in a deblurring result. We measure the blurriness of a deblurring result with the following state-of-the-art sharpness measures from recent work:

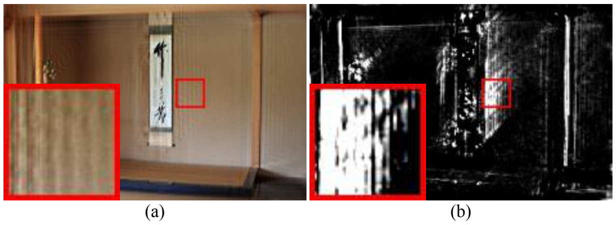

__Figure 5:__ Measuring ringing artifacts. (a) Deblurred image. (b) Ringing feature response map. For visualization, we add up responses over all pyramid levels, and convert it into a grayscale image where pixel intensities indicate the strength of the response. Original image courtesy Tanaka Juuyoh.

- **Autocorrelation (AutoCorr)** of image derivatives can be an effective way to measure the remaining blur, since it is close to the autocorrelation of the blur kernel due to the power law on the power spectrum of natural images [Field and Brady 1997; Goldstein and Fattal 2012]. Unfortunately, it is known to be vulnerable to long straight edges. We detect long straight edges using the Hough transform, and mask them out from image derivatives before computing autocorrelation. We measure how much the autocorrelation map is spread out and use it as a sharpness measure. Please refer to the supplementary material for details.
- **Cumulative Probability of Blur Detection (CPBD)** is a no-reference sharpness measure proposed by [Narvekar and Karam 2011]. It defines sharp edges based on the notion of "just noticeable blur", and measures the proportion of sharp edges as a sharpness measure.
- **Local Phase Coherence (LPC)** [Hassen et al. 2010] is a no-reference sharpness measure based on the coherence of local phase information across different scales.
- **Normalized Sparsity Measure (NormSps)** [Krishnan et al. 2011] is a measure that favors sharp images to blurry ones, and was originally proposed for blind deconvolution.

Note that some previous approaches [Samadani et al. 2010; Trentacoste et al. 2011] also measure sharpness. However, Samadani et al. [2010] use the shape of the PSF as its prior knowledge, and only Gaussian function is tested. Similarly, Trentacoste et al. [2011] make a strong assumption that the PSF is a Gaussian function. In our application, we do not know the shape of the PSF, which could be arbitrary. The PSFs estimated by different delurring algorithms from the same input image are usually different, thus none of their shapes can be trusted. Therefore, previous methods assuming known (typically Gaussian) PSF shape are not suitable for our application.

5 The Perceptually-validated Metric

In this section, we provide performance analysis on the features defined in Sec. 4, and show that any single feature cannot cover all kinds of artifacts that can appear in deblurring results. Then, we train a perceptually-validated metric based on the user study results (Sec. 3) and a combination of features, and provide its performance analysis. For performance analysis on each feature and our trained metric, we first begin by describing our evaluation method.

5.1 Evaluation Method

To evaluate a feature and our metric, we adopt the idea of ranking comparisons to evaluate how well a feature or a metric can distinguish the relative quality difference between different deblurring results. Specifically, we compute the feature/metric scores f for all images in the data set, and then rank the images based on f. We then compare this ranking with the "ground truth" score δ. δ is generated by fitting a Bradley-Terry model to the user data (Sec. 3.3) for all data sets in our experiment, except for the synthetic single-distortion image data sets (Sec. 5.2). If these two rankings correlate well, then the feature in question is a good stand-in for the perceived quality of a deblurred image.

Two widely used methods for comparing ranking results are Spearman's rank correlation [Spearman 1904] and Kendall τ distance [Kendall 1938]. However, these methods are not suitable in our application, because they do not consider two important factors. First, the distance between two adjacent images in the ground truth ranking is not uniform. Therefore, reversing a pair with a larger distance is worse than reversing a pair having a smaller distance. Second, accurate ranking among relatively good results is more important than ranking among bad ones for real applications, because bad deblurring results are typically so bad that the exact ranking among them is meaningless.

We thus propose a new ranking metric, named *weighted Kendall τ distance*, for performance evaluation. We first define a set $D_{(\delta,f)}$ of pairs (i,j), whose orders by δ and f do not agree, i.e., $(\delta_i - \delta_j)(f_i - f_j) < 0$. Kendall τ distance is defined as the cardinality of $D_{(\delta,f)}$. As this distance is solely defined based on the cardinality, it has the problems mentioned above. To overcome those problems, we define a weighted distance as:

$$M(\delta, f) = \sum_{(i,j)\in D_{(\delta,f)}} \left|\left(\max(\delta_i,\delta_j) - \delta_{\min}\right)(\delta_i - \delta_j)\right|, \quad (6)$$

where $\delta_{\min}$ is the worst B-T score. For comparison with other features whose scores have different scales, we use a normalized version of Eq. (6), which is defined as $\overline{M}(\delta, f) = M(\delta, f)/M(\delta, -\delta)$, where $M(\delta, -\delta)$ is the maximum mismatch generated by comparing against the exact opposite ranking.

The two factors in Eq. (6) are the key difference between weighted Kendall τ distance and the conventional Kendall τ distance. The first factor $(\max(\delta_i, \delta_j) - \delta_{\min})$ in Eq. (6) is larger for the highly-ranked images in the B-T model, placing an emphasis on them relative to images with worse rankings. The second factor $(\delta_i - \delta_j)$ measures mis-ranking between the target feature and the B-T model. Note that if we simply add up 1 for all pairs (i,j) in $D_{(\delta,f)}$ instead of these two factors, Eq. (6) becomes the Kendall τ distance.

5.2 Evaluating Features on Single-Distortion Images

Based on the evaluation method proposed above, we first evaluate the usefulness of each feature on different kinds of artifacts. We design three new data sets. Each one contains only one artifact among noise, blur and ringing.

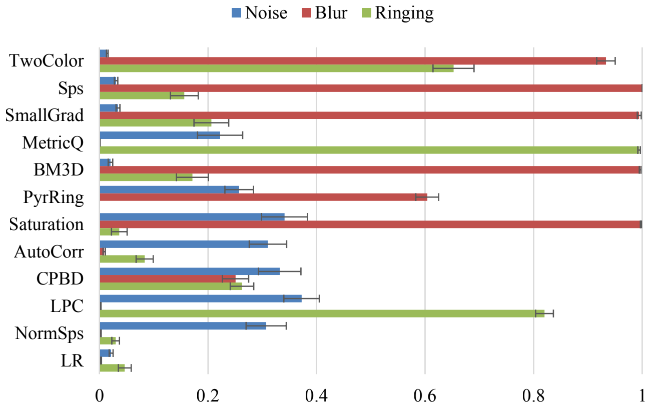

Figure 6: *The mean weighted Kendall τ distance from each individual feature and our LR metric to the ground truth. Lower is better. The error bars indicate the standard error of the distance.*

- **Noise:** The noise data set consists of deblurred images with different noise levels. To generate deblurred images, we generated synthetically blurred images with different amounts of Gaussian noise, and applied different deblurring algorithms. The standard deviation of Gaussian noise varies from 0 to 0.04 with the step 0.004 (assuming that image intensities are normalized into $[0,1]$). We used 16 original sharp images and five different deblurring algorithms to build a data set consisting of $16 \times 5 = 80$ data groups.
- **Blur:** We add synthetic motion blur with PSFs with the same pattern but different sizes onto a sharp image to build a data group. Specifically, each PSF is resized with ratio from 0.25 to 2.25 with the step 0.25. The images blurred with larger PSFs are more blurry. We use 16 sharp images, and eight original PSFs to build a data set consisting of $16 \times 8 = 128$ data groups.
- **Ringing:** Inaccurate PSFs often cause ringing artifacts. Thus, we generate a series of inaccurate PSFs by upsampling the true PSF. We found higher upsampling ratio yields more severe the ringing artifacts. Therefore, in each data group, we use a non-blind deblurring algorithm with Gaussian derivative prior to deblur images using increasingly upsampled versions of the true PSF. We use 16 sharp images and eight original PSFs to build a data set consisting of $16 \times 8 = 128$ data groups.

Since we intentionally create images with increasing amounts of artifacts in each group, here the ground truth score of the i-th image δ_i is specially defined as $\delta_i = i$ instead of conducting a user study.

We measure the quality of our features on these three data sets. Fig. 6 shows the mean weighted Kendall τ distance of each individual feature. We make the following observations:

- All noise features except MetricQ mis-classify blurry images as good images.
- Sharpness features work moderately well on the noise data set. However, MetricQ and LPC mis-classify images with ringing artifacts as good images.
- PyrRing works well on images with ringing artifacts. Saturation, AutoCorr, and NormSps also have good performance on the ringing data set.

In summary, all of these features perform quite well on at least one data set. However, none of them are able to work on all the three data sets just by themselves, motivating our approach (described in the next section) of using learning algorithms to combine features. We refer the readers to the supplementary materials for additional evaluation of each feature on our user study data set.

5.3 Learning a Metric

To derive a quality metric, we first define a metric as a mapping function from a feature space $\mathbf{X}$ to a scalar in $[0, \infty)$. The feature space $\mathbf{X}$ is defined as the concatenation of all features defined earlier. Statistical regression methods are used to fit the mapping function $f(\mathbf{x})$ based the user study results.

In this work we use a commonly-used regression method: Logistic Regression (LR) [Hilbe 2009]. LR is a well-known generalized linear regression method, which is also used for deriving scores for the Bradley-Terry model. Similarly to Sec. 3.3, suppose we have a pair of images (A, B), and there are n user study submissions for this pair with a submissions favoring image A over B and b submissions favoring B over A. As in Sec. 3.3, we can define the following logistic regression model:

$$p_{AB} = \text{logit}^{-1}(\boldsymbol{\gamma} \cdot (\mathbf{x}_A - \mathbf{x}_B)), \tag{7}$$

where $\mathbf{x}_A$ and $\mathbf{x}_B$ are feature vectors of images A and B, respectively, and $\boldsymbol{\gamma}$ is the parameter vector for logistic regression. Then, following a similar approach to Sec. 3.3, we can derive the likelihoods for all pairs (A, B), which have the same form as Eq. (5). The derived likelihoods are functions of $\boldsymbol{\gamma}$, and we can solve for $\boldsymbol{\gamma}$ with Maximum Likelihood Estimation.

Once $\boldsymbol{\gamma}$ is obtained, a metric $f(\mathbf{x})$ can be derived as follows:

$$f(\mathbf{x}) = \boldsymbol{\gamma} \cdot (\mathbf{x} - \mathbf{x}_O), \tag{8}$$

where $\mathbf{x}_O$ is the feature vector of the 'ideal' sharp image. Since $\boldsymbol{\gamma} \cdot \mathbf{x}_O$ is a constant, it can be omitted and the final form of $f(\mathbf{x})$ becomes:

$$f(\mathbf{x}) = \boldsymbol{\gamma} \cdot \mathbf{x}. \tag{9}$$

5.4 Feature Selection

Our collection of features described in Sec. 4 contains 11 different features. To train a perceptually-validated metric, the straightforward solution is to use all of the features, but this increases the chance of overfitting. In addition, features have redundancy between each other. We therefore design the following four cross-validation tests in order to select the optimal subset of features:

1. Divide the 40 data groups into five sets, in which each has eight data groups. For each set, use all images in this set for testing, and all images in all other data groups for training. By rotating the testing set we get five cross-validation errors. The mean error is then recorded as the average performance.

2. Similar as Test 1, but only use images blurred by the first PSF for training, and images blurred by the second PSF for testing, and measure the cross-validation error.

3. Similar as Test 1, but only use images with the noise levels 0 and 0.02 for training, and images with the noise level 0.01 for testing, and measure the cross-validation error.

4. Similar as Test 1, but only use images with the first PSF and noise levels 0 and 0.02 for training, and images with the second PSF and noise level 0.01 for testing, and measure the cross-validation error.

Features	Sps	SmallGrad	MetricQ
Scaled Weights	0.7344	0.1774	0.4106
Features	NormSps	AutoCorr	CPBD
Scaled Weights	0.7998	1.9179	0.4722
Features	PyrRing	Saturation	
Scaled Weights	1.7671	0.2283	

Table 1: *Scaled weights of the selected features.*

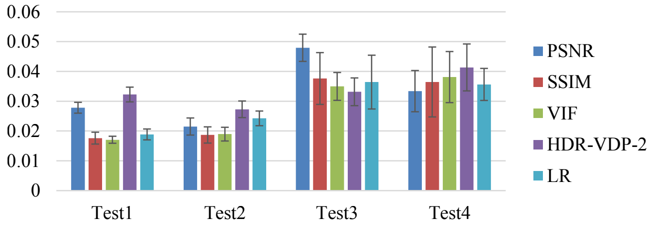

Figure 7: *The weighted Kendall τ distance of different metrics on four different cross-validation test sets (see Sec. 5.4). The error bars indicate the standard error of the distance.*

In this way, Tests 2–4 perform validation not only on image sets that were omitted from training, but also on PSFs (Tests 2 and 4) and noise levels (Tests 3 and 4) that were not used during training. This provides a better measure of generalization than cross-validation over images alone, as in Test 1.

Since we have 11 different features, there are only $2^{11} - 1 = 2047$ possible combinations of features. Therefore, we exhaustively search through all possible combinations, and select the one that generates the best average performance of the four tests.

In the end, this method selects the following features: *Sps*, *SmallGrad*, *MetricQ*; *PyrRing*, *Saturation*; *AutoCorr*, *CPBD*, *NormSps*. Table 1 demonstrates the weights of these features in our metric. Here the weights are scaled with the standard deviation of the features to reveal the importance of each feature. We have the following observations: First, all three types of single-distortion artifacts are included with non-trivial weights, which confirms our observation that they are all relevant to deblurred image quality. Second, the sharpness features have the highest overall weight, which is reasonable because this metric is for deblurring. Third, ringing features have a higher overall weight than noise features, which confirms our observation that users are sensitive to ringing artifacts.

Fig. 6 demonstrates that our LR metric performs very well on all three types of data sets. Fig. 7 compares the performance of our metrics trained with the optimal subset of features and three existing full-reference metrics (PSNR, multi-scale SSIM, VIF, and HDR-VDP-2), using the evaluation method described in Sec. 5.1. Our LR metric is competitive with all full-reference metrics. Given that our metric is not only competitive with the existing ones but is also no-reference (i.e., does not require ground truth), we believe that its performance is promising for real-world applications, as shown in Sec. 6.

5.5 Validation of User Study on Mechanical Turk

Because users on Amazon Mechanical Turk are unsupervised and come from all over the world, the variance of data can be considerably larger than in controlled in-lab psychophysical experiments. We therefore analyze the repeatability of our user study.

There are several ways to evaluate the consistency of a user study. A common approach might be to look at inter-subject variance. However, since paired comparison actually expects (and indeed relies on) disagreement on pairs of images that have similar quality, inter-observer variance is not applicable for this methodology.

To avoid this problem, we instead measure the inter-phase variance of our user study experiments. We repeat user study experiments of five out of the 40 data groups on Amazon Mechanical Turk, three months after the original experiments, and fit a new score δ' for each image in the data groups.

- Consider the original δ as the reference. The weighted Kendall τ distances of the new δ' of the five data groups are all below 1.21×10^{-3}.
- Consider the new δ' as the reference. The weighted Kendall τ distances of the original δ of five data groups are all below 1.35×10^{-3}.

These two results prove that our user study results are stable.

6 Applications and Results

We now demonstrate how our deblurring quality metric can be applied in different application scenarios.

6.1 Automatic Parameter Selection

In previous work image quality metrics have been applied to automatically select good parameter settings for image processing algorithms, such as image denoising [Zhu and Milanfar 2010; Mittal et al. 2012a]. We demonstrate that our metric is helpful for parameter selection in a new problem: motion deblurring. We observe that there are two important parameters that almost all deblurring algorithms require: (1) the regularization strength for the final non-blind deconvolution step, which controls the trade-off between noise and residual blur in the final result; and (2) the maximum PSF size. We found that existing deblurring systems are sensitive to these two parameters, and there is no principled way to find good settings for them. Thus, previous deblurring methods often require extensive manual parameter tuning to generate good results.

To demonstrate that our metric can be used for automatic parameter selection, we conduct an experiment with six synthetic examples that are not included in our user study data set. The images are blurred with two PSFs that are also different from those used in the study. Gaussian noise with $\sigma = 0.01$ is added to each test image to form a parameter selection dataset. For this study we choose two algorithms:

1. The fast PSF estimation approach proposed by Cho and Lee [2009] to estimate PSFs, and a non-blind deblurring method with a Gaussian prior for generating the final result, which is formulated as:

$$\operatorname*{argmin}_{l} \left\{ \|b - l * k\|^2 + \lambda_C \left(\|f_x * l\|^2 + \|f_y * l\|^2 \right) \right\}, \qquad (10)$$

where the first term is a data fidelity term, and the second term is a regularization term. λ_C is a regularization weight, and f_x and f_y are first-order derivative filters along the x and y directions.

2. The PSF estimation approach proposed by Levin et al. [2011], and with the non-blind deblurring method with a sparse derivative prior proposed by them:

$$\operatorname*{argmin}_{l} \left\{ \|b - l * k\|^2 + \lambda_L \left(\sum_{i=1}^{2} |f_i * l|^{\alpha} + 0.25 \sum_{i=1}^{3} |g_i * l|^{\alpha} \right) \right\}, \qquad (11)$$

where the f_1 and f_2 are first-order derivative filters, and g_1, g_2, and g_3 are second-order derivative filters. α is 0.8 to impose sparsity.

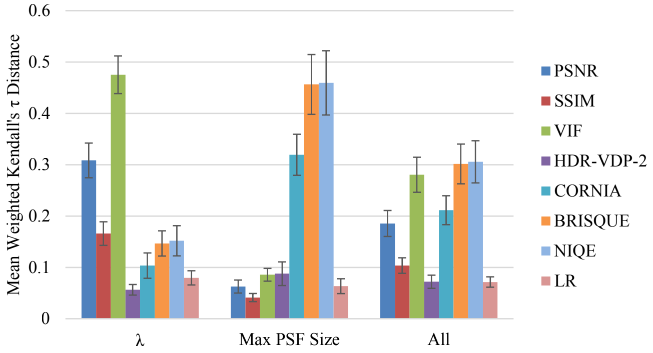

Figure 8: *The mean weighted Kendall τ distance of full-reference (PSNR, SSIM, VIF, HDR-VDP-2) and no-reference (CORNIA, NIQE, BRISQUE, LR) metrics against the Bradley-Terry model on the data set of automatic parameter selection. Lower values indicate better performance.*

To select the PSF size, we try a sequence of different sizes from 11×11 to 61×61 with a step size of 5, while fixing the regularization strength $\lambda_C = 2^{-3}$ and $\lambda_L = 2^{-8}$, and use our metric to find the PSF size which results in the highest score. Secondly, we fix the selected PSF size, and apply a sequence of different λ_C values: $\{2^{-9}, 2^{-8}, \ldots, 2^{3}\}$ and λ_L values: $\{2^{-14}, 2^{-13}, \cdots, 2^{-2}\}$, and use our metric again to find the optimal value for λ.

To evaluate the parameter selection results, we conduct another pair-by-pair user study on the parameter selection data set, where each pair was ranked by at least 20 users. The Bradley-Terry model is then used to get a ground truth score for each image in this data set. We compare our Logistic Regression (LR) metrics with existing full-reference (PSNR, SSIM, VIF, and HDR-VDP-2), and a few state-of-the-art general-purpose no-reference image quality metrics, including CORNIA [Ye et al. 2012], BRISQUE [Mittal et al. 2012b] and NIQE [Mittal et al. 2012c]. Fig. 8 shows the mismatch scores of each method.

The following conclusions can be drawn from this experiment:

- For tuning λ, HDR-VDP-2 and our LR have the best performance. SSIM and the three no-reference metrics work reasonably well. PSNR and VIF perform worst.
- For tuning the maximum PSF size, all full-reference metrics and our LR have very good performance. However, the other no-reference metrics perform badly.

Overall, our LR achieves the best mean Kendall τ distance. We run a one-tailed paired t-test with confidence level 0.95 to compare our method against each other method to check the statistical significance of its superiority. The p-value is below 0.05 for each metric except HDR-VDP-2, which suggests that our LR metric is statistically significantly better. The p-value for HDR-VDP-2 is 0.4864, which indicates that our LR metric is statistically comparative with it. Note that HDR-VDP-2 is a *full-reference* metric, while ours is *no-reference*.

In order to better understand why our LR metric outperforms other metrics in general but not in λ tuning, we provide the following case studies. The full set of images and their rankings can be found in the supplementary materials.

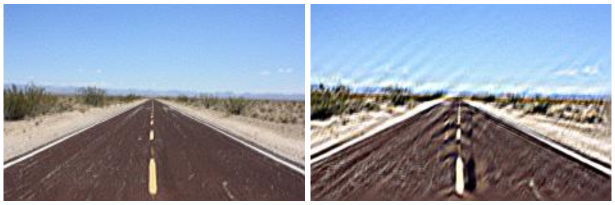

Figure 9: *Deblurring results with max PSF size of 21 and 61. Original image courtesy Semio@Flickr.*

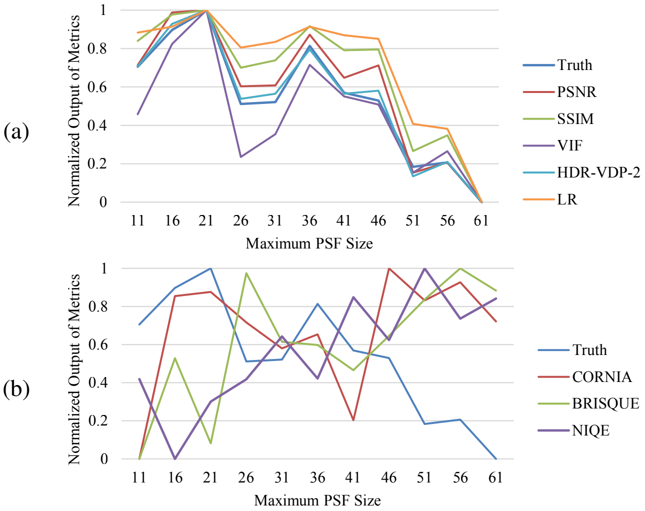

Figure 10: *Comparing performance of different metrics in Case 1. Closer to the ground truth curve means better performance.*

Case study 1. The test image "road" is challenging for PSF estimation, because there are too few strong edges that PSF estimation methods typically rely on. Inaccurate PSF estimation will in turn introduce strong ringing artifacts into the deblurring results, as shown in Fig. 9. In particular, more severe ringing artifacts can be observed with larger PSF sizes. In Fig. 10 we compare the full-reference and no-reference metrics with the ground truth, which is obtained by the Bradley-Terry model. Here we normalize the outputs of all metrics to make them lie in $[0, 1]$. It shows that the full-reference metrics and our LR metric yield nearly the same trend as the ground truth, indicating a correct evaluation of the perceptual quality of the results. In contrast, the results of other no-reference metrics such as CORNIA, NIQE, and BRISQUE present large divergence from the ground truth, indicating poor performance. This is because the artifacts in deblurred images, particularly the residual motion blur and the very severe ringing artifacts, are quite different from artifacts in general-purpose image quality assessment data sets, thus the no-reference metrics trained from general data sets cannot work well on deblurred images.

Case study 2. On the other hand, when it comes to the trade-off between a little amount of residual blur and noise (i.e. the selection of λ), the problem is quite similar to general-purpose image quality assessment, as there are very few deblurring-specific artifacts involved. In this case, the general no-reference metrics perform much better, as shown in the comparison in Fig. 11. The peak of the truth curve indicates perceptually the best result. Images to its left are too noisy, and those to its right are too blurry. It shows that NIQE and our LR metric best match the left part of the truth curve; however, for the right part of the truth curve, we observe that all metrics favor blurry images slightly too much, including our metric.

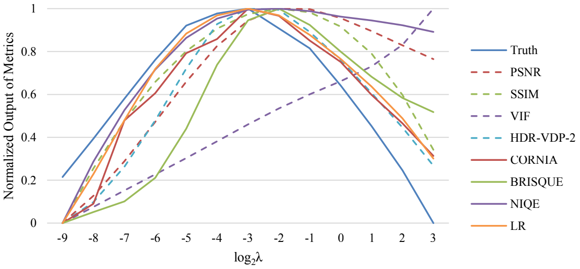

Figure 11: *Comparing performance of different metrics in Case 2.*

6.2 Algorithm Selection

Given a blurry input image, we can apply different algorithms to deblur it, and then use our metric to automatically choose the best result. To demonstrate this we test our metric on the recently constructed motion blur data set proposed by Köhler et al. [2012], which contains real motion-blurred images and their corresponding sharp latent images, as well as deblurring results from eight different algorithms. There are four scenes in this data set, and each is blurred with 12 camera motion trajectories controlled by a robot. The 3rd trajectory is trivial, and all algorithms achieve good results. The 8th, 9th, and 10th trajectories are so large that no algorithm can generate reasonable results. We omit these four trajectories. We also omit the results of [Fergus et al. 2006], because they consistently contain saturated pixel values in the blue channel, possibly due to an improper parameter setting. In total we thus have 32 data groups, each deblurred by seven different algorithms.

We applied our metric to rank the images in each data group. To evaluate the results, we conduct another pair-by-pair user study on this data set, where each pair was ranked by at least 20 users. The Bradley-Terry model is then used to get a ground truth score for each image in this data set. Fig. 12 shows the mean weighted Kendall τ distance from all metrics to the Bradley-Terry model. For this application, our LR metric performs slightly better than all full-reference metrics except SSIM, and significantly better than all the other no-reference metrics. We run a one-tailed paired t-test with confidence level 0.95 to compare our method against each other method. The p-values are 0.1159, 0.0340, and 0.1068 when we compare LR to the full-reference metrics PSNR, VIF, and HDR-VDP-2. The p-values are all below 10^{-4} when we compare LR to all other no-reference metrics. In Fig. 13 we show the result of one data group. The ranking generated by our LR metric matches well with the ranking given by the Bradley-Terry model, particularly for top three images.

6.3 Image Fusion

In addition to selecting the single best-performing algorithm for an input image, we can go further and use our metric to select the best local regions from multiple deblurred images and fuse them into a deblurred image that is better than any one result. This can also be used to handle spatially-varying blur. Specifically, we could estimate multiple PSFs from different local regions of an input image, and use them to generate multiple deblurring results. Our metric is then used to automatically find good regions in different versions and stitch them together to form the best possible result.

We conduct an experiment to validate this idea. First, we create several images with spatially-varying blur, using the tools provided by Whyte et al. [2010], and add Gaussian noise with $\sigma = 0.001$ to

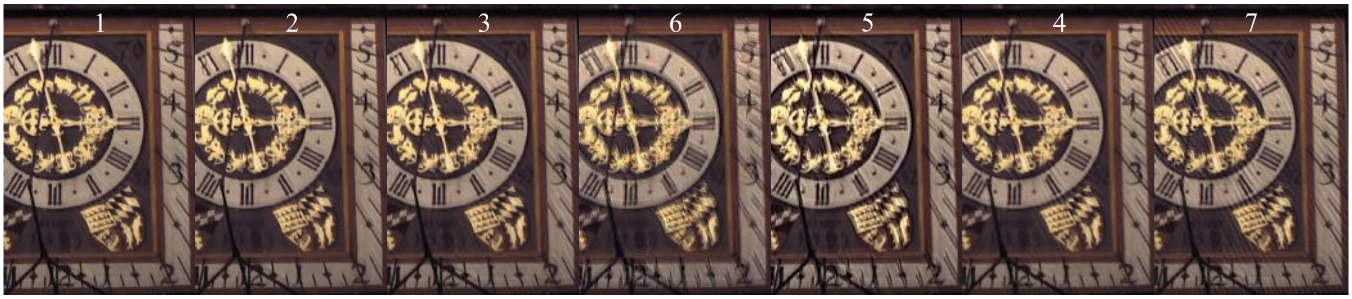

Figure 13: *Using our metric for algorithm selection. The images are sorted by our LR metric with decreasing quality from left to right. The white numbers on top of images are the ranking given by the Bradley-Terry model.*

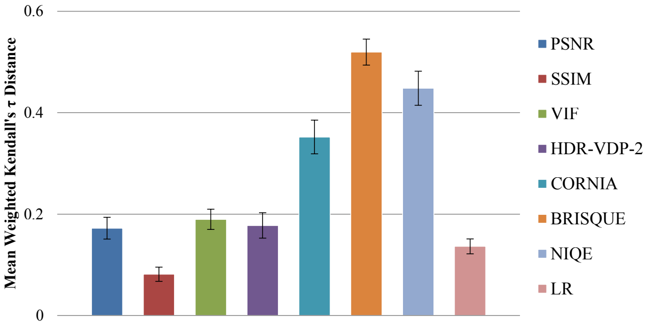

Figure 12: *The mean weighted Kendall τ distance of full-reference (PSNR, SSIM, VIF, HDR-VDP-2) and no-reference (CORNIA, NIQE, BRISQUE, LR) metrics against the Bradley-Terry model on the data set of algorithm selection. Lower values indicate better performance.*

them. Fig. 14(a) shows one of these examples. We design a simple algorithm to deblur these images. First, we divide the image into a 3×3 grid of regions, and estimate a blur kernel in each region using the kernel estimation method in [Cho and Lee 2009]. We then use the non-blind deconvolution method from the same approach to recover a latent image from each kernel, resulting in nine deblurred images in total. Each deblurred image has some good regions, but also has severe artifacts in other areas.

To fuse these results, we first align all images using normalized cross correlation. We then apply our metric on small regions surrounding each pixel, yielding an evaluation score for each pixel in each image. We divide each image into overlapping patches with 11×11 pixels, and for each patch, we select the one from all results with the highest mean metric values. The selected patches are stitched together using graph cuts and Poisson blending [Pérez et al.] to generate the fused result.

Fig. 1 and Fig. 14 show two examples of fusing multiple deblurred images. We compare our results with two methods. In the first method, we compute the average of all images, which is helpful for reducing artifacts that are uncorrelated in each deblurred image. For the second method, in each deblurred image we pick the same region that the algorithm uses to estimate the blur kernel, and combine them to obtain a naive fusion result. As shown in these examples, our results contain fewer artifacts and have better visual quality compared with results generated from these methods, as well as the individual images. Fig. 15 shows our fusion result on a real-world spatially-varying blurry photo from [Whyte et al. 2010]. Our result is comparable to the result of the spatially varying deblurring algorithm [Whyte et al. 2010]. Note that in this test case [Whyte et al. 2010] uses an additional sharp noisy image as auxiliary information, while ours does not.

For a better comparison, we conduct a user study on a collection of 19 synthetic images with spatially-varying blur. Users were asked to do pairwise comparison as we described in Sec. 3. Each pair of images is compared at least 37 times. The proportion of users who favor our fusion results over that of the simple averaging method and the naive fusion method are $802/828$ (96.86%) and $644/838$ (76.85%), respectively. The proportion of users who favor the naive fusion results over the simple averaging results is $724/818$ (88.51%). A randomized permutation test on the distributions of users' preferences shows the statistical significance ($p \ll 0.01$) of all these results. Please refer to the supplementary materials for the collection of images used in our user study.

7 Conclusion and Future Work

We have demonstrated a perceptually-validated no-reference metric for evaluating the quality of image deblurring results. To achieve this we conduct a user study to collect users' evaluations on the visual quality of deblurred images. By studying user data we identify the three most common deblurring artifacts, and design a collection of features for measuring them. We further show how to select an optimal subset of features and use them to train a metric. Extensive evaluation shows that this outperforms state-of-the-art no-reference metrics, and matches or outperforms full-reference metrics, for evaluating deblurred images. Finally, we demonstrate that our metric can be used for improving image deblurring by parameter selection, algorithm selection, and fusion of multiple results.

As discussed in Sec. 6.1, our metric currently favors slightly-blurry results, a common problem for all metrics we have tested. As future work, we plan to explore how to remove this bias from the metric. We would also like to explore how to design better image deblurring algorithms that explicitly maximize the evaluation score of the metric to generate results with high visual quality.

Acknowledgements

We would like to thank Sylvain Paris, Rebecca Fiebrink, Su Xue, and the Princeton University Graphics Group for their helpful discussion, and the SIGGRAPH Asia reviewers for their comments. We also thank various Flickr users for sharing their photos, and the authors of previous work for sharing their code and data. This work is supported by NSF grants CCF-1012147, CCF-1027962 and a gift from Adobe.

References

BRADLEY, R. A., AND TERRY, M. E. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons.

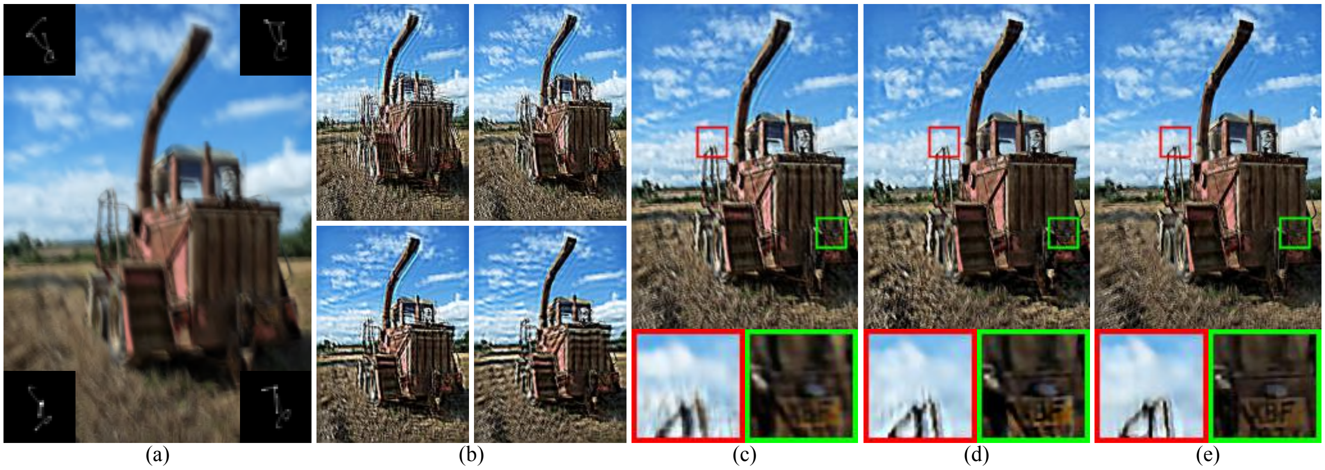

Figure 14: *Using our metric to fuse multiple deblurring results from an image which contains spatially-varying blur. (a) The input image with spatially varying blur, as well as the PSFs at the four corners. (b) Four out of the nine deblurring results produced with the deblurring algorithm in [Cho and Lee 2009] that has the assumption of spatially invariant blur. (c) the result by simply averaging all deblurring results. (d) result by the naive fusion method (see Sec. 6.3). (e) our fusion result. Original image courtesy Alex Brown.*

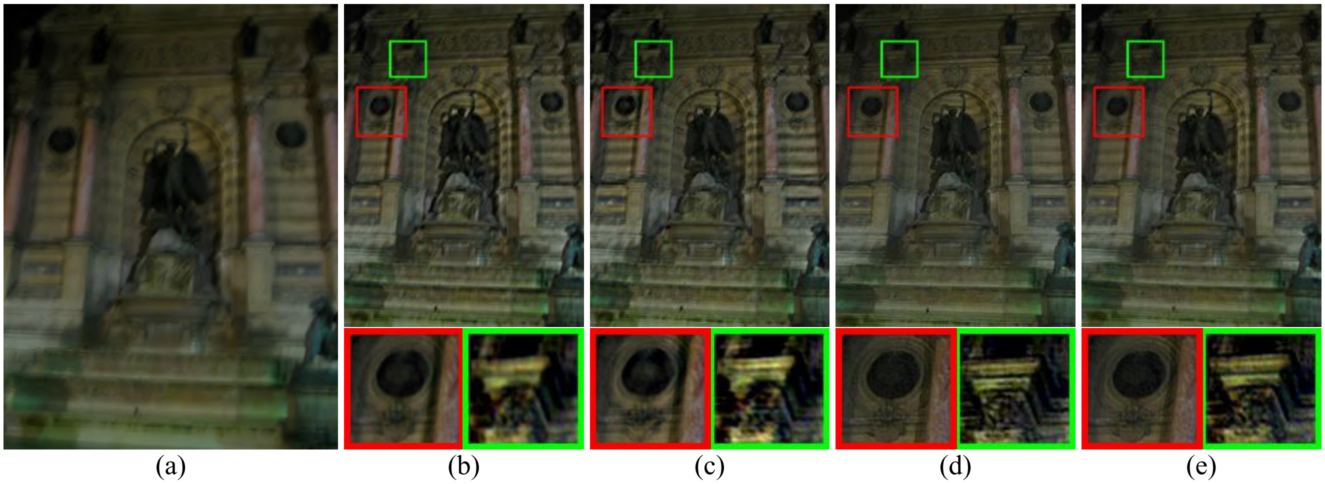

Figure 15: *Using our metric to fuse multiple deblurring results of a real-captured spatially-varying blurry photo from [Whyte et al. 2010]. (a) The input blurry image. (b) the result by simply averaging all deblurring results. (c) result by the naive fusion method. (d) result of [Whyte et al. 2010]. (e) our fusion result.*

Biometrika 39, 3/4.

CADIK, M., HERZOG, R., MANTIUK, R., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2012. New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts. *ACM Trans. Graphics 31*, 6 (Nov.).

CHEN, X., GOLOVINSKIY, A., AND FUNKHOUSER, T. 2009. A benchmark for 3D mesh segmentation. *ACM Trans. Graphics 28*, 3.

CHO, S., AND LEE, S. 2009. Fast motion deblurring. *ACM Trans. Graphics 28*, 5.

CHO, T. S., PARIS, S., HORN, B., AND FREEMAN, W. 2011. Blur kernel estimation using the Radon transform. In *Proc. CVPR 2011*.

COLE, F., SANIK, K., DECARLO, D., FINKELSTEIN, A., FUNKHOUSER, T., RUSINKIEWICZ, S., AND SINGH, M. 2009. How well do line drawings depict shape? *ACM Trans. Graphics 28*, 3.

DABOV, K., FOI, A., KATKOVNIK, V., AND EGIAZARIAN, K. 2007. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Trans. Image Processing 16*, 8.

FERGUS, R., SINGH, B., HERTZMANN, A., ROWEIS, S. T., AND FREEMAN, W. T. 2006. Removing camera shake from a single photograph. *ACM Trans. Graphics 25*, 3.

FIELD, D. J., AND BRADY, N. 1997. Visual sensitivity, blur and the sources of variability in the amplitude spectra of natural scenes. *Vision Research 37*, 23.

GOLDSTEIN, A., AND FATTAL, R. 2012. Blur-kernel estimation from spectral irregularities. In *Proc. ECCV 2012*.

GUPTA, A., JOSHI, N., ZITNICK, L., COHEN, M., AND CURLESS, B. 2010. Single image deblurring using motion density functions. In *Proc. ECCV 2010*.

HASSEN, R., WANG, Z., AND SALAMA, M. 2010. No-reference image sharpness assessment based on local phase coherence measurement. In *Proc. ICASSP 2010*.

HILBE, J. M. 2009. *Logistic Regression Models*. Chapman & Hall/CRC Press.

JI, H., AND WANG, K. 2012. A two-stage approach to blind spatially-varying motion deblurring. In *Proc. CVPR 2012*.

JOSHI, N., SZELISKI, R., AND KRIEGMAN, D. 2008. PSF estimation using sharp edge prediction. In *Proc. CVPR 2008*.

JOSHI, N., ZITNICK, C., SZELISKI, R., AND KRIEGMAN, D. 2009. Image deblurring and denoising using color priors. In *Proc. CVPR 2009*.

KENDALL, M. G. 1938. A new measure of rank correlation. *Biometrika 30*, 1/2.

KÖHLER, R., HIRSCH, M., MOHLER, B., SCHÖLKOPF, B., AND HARMELING, S. 2012. Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database. In *Proc. ECCV 2012*.

KRISHNAN, D., AND FERGUS, R. 2009. Fast image deconvolution using hyper-Laplacian priors. In *Proc. NIPS 2009*.

KRISHNAN, D., TAY, T., AND FERGUS, R. 2011. Blind deconvolution using a normalized sparsity measure. In *Proc. CVPR 2011*.

LEVIN, A., FERGUS, R., DURAND, F., AND FREEMAN, W. T. 2007. Image and depth from a conventional camera with a coded aperture. *ACM Trans. Graphics 26*, 3.

LEVIN, A., WEISS, Y., DURAND, F., AND FREEMAN, W. 2011. Efficient marginal likelihood optimization in blind deconvolution. In *Proc. CVPR 2011*.

MANTIUK, R., KIM, K. J., REMPEL, A. G., AND HEIDRICH, W. 2011. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graphics 30*, 4.

MARZILIANO, P., DUFAUX, F., WINKLER, S., AND EBRAHIMI, T. 2004. Perceptual blur and ringing metrics: Application to JPEG2000. *Signal Processing: Image Communication 19*, 2.

MASIA, B., PRESA, L., CORRALES, A., AND GUTIERREZ, D. 2012. Perceptually optimized coded apertures for defocus deblurring. *Computer Graphics Forum 31*, 6.

MITTAL, A., MOORTHY, A. K., AND BOVIK, A. C. 2012. Automatic parameter prediction for image denoising algorithms using perceptual quality features. In *Proc. SPIE*, vol. 8291.

MITTAL, A., MOORTHY, A., AND BOVIK, A. 2012. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Processing 21*, 12.

MITTAL, A., SOUNDARARAJAN, R., AND BOVIK, A. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters PP*, 99.

MOORTHY, A., AND BOVIK, A. 2010. A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters 17*, 5.

NARVEKAR, N., AND KARAM, L. 2011. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). *IEEE Trans. Image Processing 20*, 9.

PÉREZ, P., GANGNET, M., AND BLAKE, A. Poisson image editing. *ACM Trans. Graphics 22*, 3.

SAAD, M., BOVIK, A., AND CHARRIER, C. 2010. A DCT statistics-based blind image quality index. *IEEE Signal Processing Letters 17*, 6.

SAMADANI, R., MAUER, T. A., BERFANGER, D. M., AND CLARK, J. H. 2010. Image thumbnails that represent blur and noise. *IEEE Trans. Image Processing 19*, 2 (Feb.).

SCHULER, C. J., HIRSCH, M., HARMELING, S., AND SCHÖLKOPF, B. 2012. Blind correction of optical aberrations. In *Proceedings of the 12th European conference on Computer Vision - Volume Part III*, Springer-Verlag, Berlin, Heidelberg, Proc. ECCV 2012, 187–200.

SECORD, A., LU, J., FINKELSTEIN, A., SINGH, M., AND NEALEN, A. 2011. Perceptual models of viewpoint preference. *ACM Trans. Graphics 30*, 5.

SHAN, Q., JIA, J., AND AGARWALA, A. 2008. High-quality motion deblurring from a single image. *ACM Trans. Graphics 27*, 3.

SHEIKH, H., AND BOVIK, A. 2006. Image information and visual quality. *IEEE Trans. Image Processing 15*, 2.

SPEARMAN, C. 1904. The proof and measurement of association between two things. *The American journal of psychology 15*, 1.

TANG, H., JOSHI, N., AND KAPOOR, A. 2011. Learning a blind measure of perceptual image quality. In *Proc. CVPR 2011*.

TEO, P., AND HEEGER, D. 1994. Perceptual image distortion. In *Proc. ICIP 1994*, vol. 2.

TRENTACOSTE, M., MANTIUK, R., AND HEIDRICH, W. 2011. Blur-Aware Image Downsizing. In *Proc. Eurographics*.

WANG, Z., BOVIK, A., SHEIKH, H., AND SIMONCELLI, E. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Processing 13*, 4.

WHYTE, O., SIVIC, J., ZISSERMAN, A., AND PONCE, J. 2010. Non-uniform deblurring for shaken images. In *Proc. CVPR 2010*.

XU, L., AND JIA, J. 2010. Two-phase kernel estimation for robust motion deblurring. In *Proc. ECCV 2010*.

YE, P., KUMAR, J., KANG, L., AND DOERMANN, D. 2012. Unsupervised feature learning framework for no-reference image quality assessment. In *Proc. CVPR 2012*.

YUAN, L., SUN, J., QUAN, L., AND SHUM, H.-Y. 2007. Image deblurring with blurred/noisy image pairs. *ACM Trans. Graphics 26*, 3.

ZHU, X., AND MILANFAR, P. 2009. A no-reference sharpness metric sensitive to blur and noise. In *Quality of Multimedia Experience 2009*.

ZHU, X., AND MILANFAR, P. 2010. Automatic parameter selection for denoising algorithms using a no-reference measure of image content. *IEEE Trans. Image Processing 19*, 12.

ZORAN, D., AND WEISS, Y. 2011. From learning models of natural image patches to whole image restoration. In *Proc. ICCV 2011*.