

Perceptually Relevant Image Fidelity

Christopher C. Taylor^a, Zygmunt Pizlo^b, and Jan P. Allebach^a

^aElectronic Imaging Systems Laboratory
Purdue University, School of Electrical and Computer Engineering
West Lafayette, IN 47907-1285
{taylor,allebach}@ecn.purdue.edu
www.ecn.purdue.edu/EISL/

^bPurdue University, Department of Psychological Sciences
West Lafayette, IN 47907-1364
pizlo@psych.purdue.edu

ABSTRACT

In recent years a number of image fidelity measures have been developed. These measures are designed to predict a person's ability to perceive differences between two nearly identical images. Successful image fidelity measures allow digital imaging developers to replace difficult and time consuming subjective evaluations with automated evaluations. Although a number of image fidelity measures have been developed, no method for evaluating and comparing the accuracy of these measures has been commonly accepted. In this paper we describe a new method for evaluating image fidelity measures. The method involves comparing spatially localized ratings from a human subject with distortion maps generated by an image fidelity measure.

Keywords: Image fidelity, image quality, evaluation, human visual system, contrast discrimination

1. INTRODUCTION

Digital imaging methods and rendering devices often introduce unwanted changes in the appearance of the images being processed. For example, lossy image compression algorithms seek to maximize compression while minimizing the amount of perceived distortions introduced to the compressed image. The compression ratio is easy to express numerically by taking the ratio of the amount of memory needed to store the original image to the amount of memory needed to store the compressed image. A numeric measure for perceived image fidelity is more elusive.

One simplistic way to estimate image fidelity is to compute the mean square error (MSE). Mean square error is the average of the squared difference between the intensity of the original and distorted images at each pixel location. It is well established, however, that MSE does not provide an adequate measure of perceived image fidelity. This is illustrated in Fig. 1. The images in Fig. 1(b) and (c) are distorted images of the original image in Fig. 1(a). The image in Fig. 1(b) has been distorted by JPEG quantization, and the image in Fig. 1(c) has been distorted by a slight shift in luminance values. The image distorted by JPEG quantization appears to be significantly

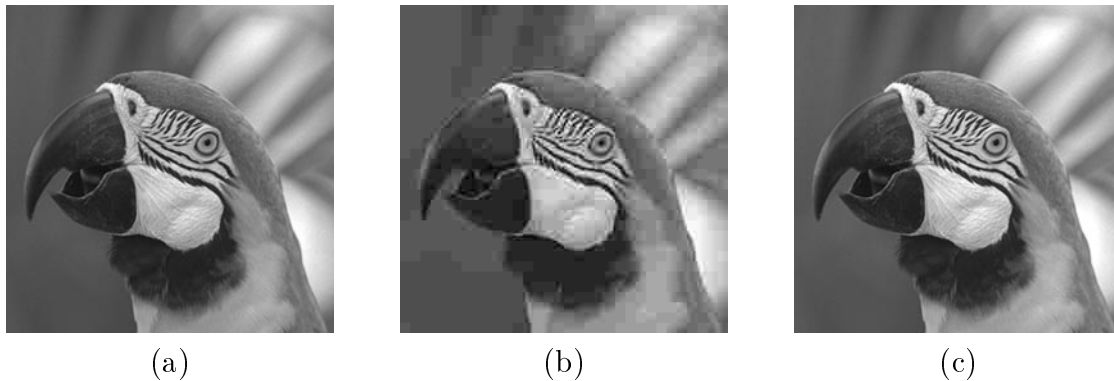


Figure 1. The perceived fidelity between the original image (a) and the JPEG quantized image (b) is much lower than the perceived fidelity between image (a) and the image with a luminance shift (c) even though the MSE between images (a) and (b) is less than the MSE between images (a) and (c).

distorted while the image distorted by a shift in luminance does not appear to suffer from significant distortion. However, the MSE for the image distorted by a shift in luminance is greater than the MSE for the image distorted by JPEG quantization. Clearly, MSE does not provide an adequate measure of perceived image fidelity.

A number of researchers have developed image fidelity measures designed to match human perception.¹⁻⁴ All these measures accept two images as input and produce a distortion map as output. The distortion map indicates, at each spatial location, the likelihood of perceiving a difference between the two input images. An example of such a map is shown in Fig. 2.

A few methods to evaluate the performance of these models have been developed. Psychophysical methods that can be used with simple stimuli have been well established, and both the Visible Differences Predictor (VDP)⁵ and the Sarnoff Visual Discrimination Model (VDM)⁶ have been compared to a wide variety of psychophysical experiments involving simple stimuli. For example, Lubin⁶ compared results from the Sarnoff VDM to a study by Blackwell and Blackwell⁷ on the detection of small luminous disks. The detection data in the psychophysical experiment consisted of visibility thresholds for luminous disks of varying disk diameter. To obtain the model's prediction, two simple images, a uniform gray image and a uniform gray image with a central luminous disk, were inputs for the Sarnoff VDM. The model produced a distortion map which was transformed into a single threshold value. By varying the luminance and diameter of the disk in the input image, the Sarnoff VDM generated visibility thresholds similar to the psychophysical data of Blackwell and Blackwell. These studies provide an important first step in validating image fidelity measures. However, these measures have been designed to be used with complex stimuli. Therefore, we argue that validation techniques that do not include complex stimuli are not complete.

Complex stimuli are more difficult to use because they are multidimensional. As a result, when the subject is presented with a pair of complex images that differ in many spatial locations, it may be difficult for the subject to provide an overall estimate of perceived differences. Therefore, until recently, validating image fidelity measures on complex images was limited to informal judgments of the distortion maps generated by a particular model. Specifically, the subject was asked to compare the distortion maps produced by an image fidelity measure to the input images and decide

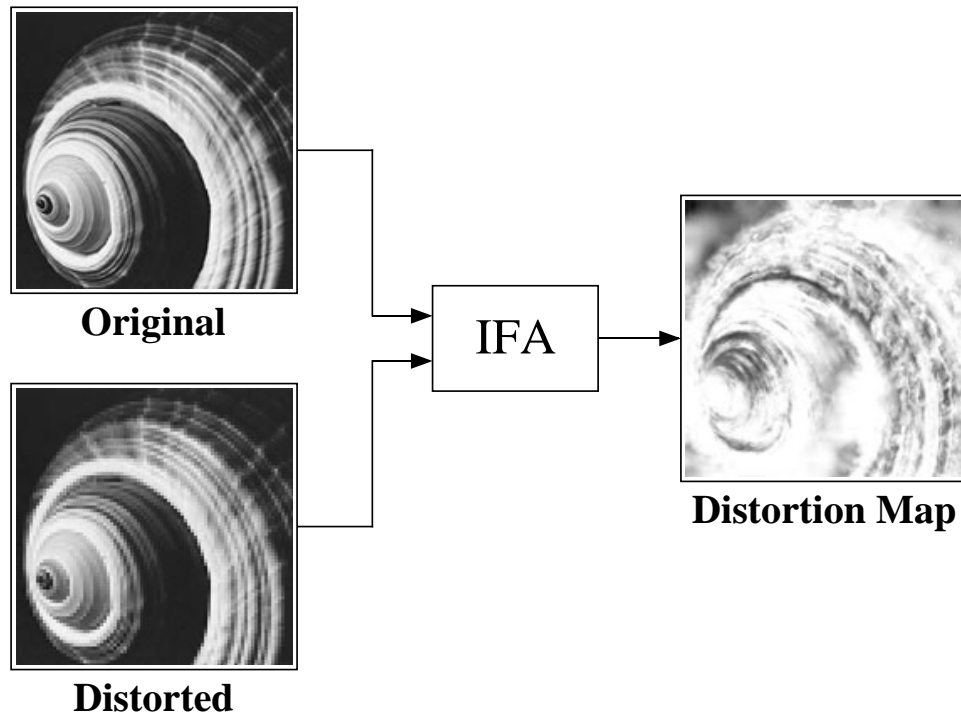


Figure 2. Example of the Image Fidelity Assessor processing. The original (upper left) and distorted (lower left) images are input into the IFA to produce the distortion map (right).

whether or not the distortion map represents visible differences between the input pair. To obtain a quantitative evaluation of models, Zhang et al.⁸ proposed a method for generating distortion maps psychophysically. In their method, subjects were presented with an original and a distorted image. Subjects used a mouse and choice of three marking sizes (circular spots with 10, 30, and 50 pixel diameters) to indicate where the image pair differed in appearance. The results from 22 subjects were pooled to produce a distortion map that indicated the probability of a subject identifying a difference between the image pair. Distortion maps like these could then be compared quantitatively with distortion maps generated by various image fidelity models. This method retains the spatial localization of the distortion maps and allows evaluation of model performance on complex images. However, the pooling stage of this method assumes that all subjects adopted the same response bias. This assumption is quite strong and is not likely to be satisfied. In fact, Zhang et al. excluded the responses of one subject from the pooling stage since the subject's response bias was unusually low, i.e., the subject "marked everywhere on every image." One further difficulty with this method is that subjects may fail to examine some portions of the image as a result of eye fixation strategy.

In the next section we describe a new method that is designed to overcome, at least partially, these difficulties. The method seeks to produce a performance metric for image fidelity measures that operates on complex stimuli, retains spatial localization, forces the subject to examine the entire image, and reduces the effects of response bias.

2. EXPERIMENT

2.1. Apparatus

Images were displayed on a calibrated 24-bit color monitor at a resolution of 100 pixels per inch on a 1024×1280 pixel screen. The peak luminance was 70 cd/m^2 and the gamma was 2.25. All images were gamma corrected and displayed in a dark room. A chin-forehead rest was used to stabilize the subject's head at a viewing distance of 0.6 m.

2.2. Subject

One of the authors (CCT) served as the subject in the experiment. He was a slight myope and wore his normal correcting glasses. The subject viewed the images binocularly with natural pupils.

2.3. Stimulus

The stimuli consisted of six original images: falls (Fig. 3), lightning (Fig. 5), parrot (Fig. 1), pattern (Fig. 6), post (Fig. 4), and shell (Fig. 2). Distorted images were created using six different types of distortion (blurring, sharpening, quantization, subsampling, etc). A total of ten image pairs were used in the experiment.

2.4. Procedure

A snapshot from an experimental trial is shown in Fig. 3. In a single trial, a distorted image and the corresponding original image were displayed side-by-side. Superimposed grid lines partitioned each image into 64 blocks. For each block the subject was asked to rate the significance of differences between the two images on an integer scale between 1 (no visible differences) and 5 (very significant visible differences). The subject selected the order in which he rated the blocks. To ensure that the ratings for each block conformed to a common scale, the subject was allowed to modify block ratings until he was satisfied with all his ratings. No time limit was set for the rating process. The subject could remove the grid lines by depressing the mouse button. The grid lines reappeared when the mouse button was released. After rating all the blocks, the subject pressed the "Done" button. This process was repeated for all ten image pairs.

2.5. Analysis

Image fidelity measures may be validated by comparing their distortion maps to the results from the psychophysical experiment. The distortion maps used in the following discussion were generated by the IFA.⁴

The nature of the results from the IFA and the psychophysical experiment differ in two ways. First, the IFA results contain a dense sampling of estimates for the perceived image fidelity across the image pair whereas the experimental results contain a sparse sampling of the perceived image fidelity across the image pair. Second, the IFA results consist of continuous probabilities whereas the experimental results consist of five discrete rating levels. The IFA results are transformed in order to remove these two differences. Direct comparisons between the IFA and experimental results can then be made.

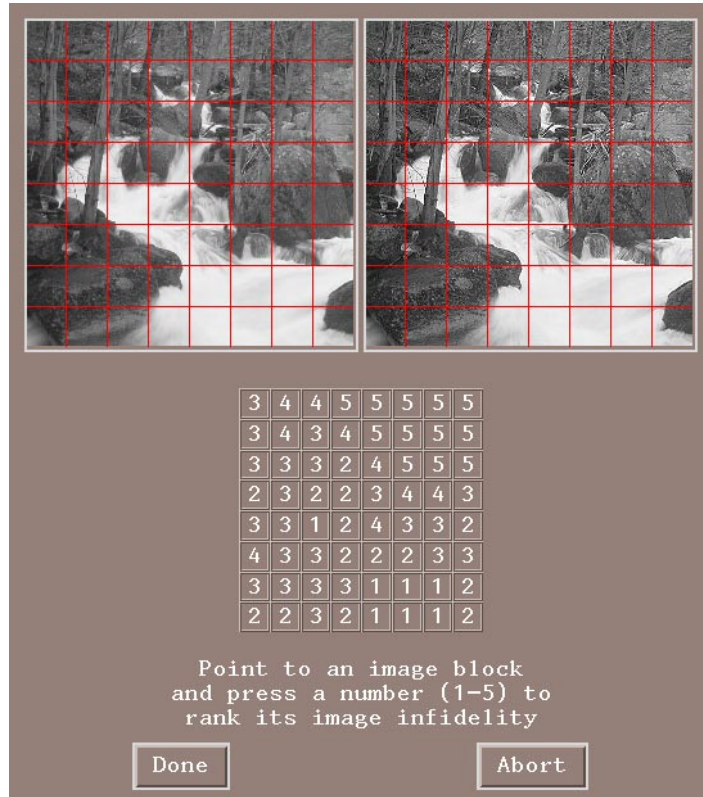


Figure 3. Snapshot of an experiment trial.

First, the densely sampled estimates of perceived image fidelity are transformed into the 8×8 arrays, based on the IFA distortion map. In one case (called the *mean IFA array*), the entries correspond to the mean of the IFA distortion map for each partitioned block. In another case (called the *maximum IFA array*), the entries correspond to the maximum of the IFA distortion map for each partitioned block.

The psychophysical rating array, mean IFA array, and maximum IFA array can be compared visually for an image pair. Two examples are shown in Figs. 4 and 5. The images in (a) and (b) are the original and distorted images. The images in (c), (d), and (e) represent the psychophysical rating array, mean IFA array, and maximum IFA array, respectively. The images in (c), (d), and (e) were generated by upsampling the arrays using pixel replication and then lowpass filtering. In both of the figures, the images associated with the psychophysical rating array and mean IFA array appear to be similar; and the maximum IFA array appears to produce overly sensitive results.

Second, the IFA predictions are converted to an 8×8 array of rating. This is done in two ways: bi-level and multi-level. In the bi-level case the psychophysical ratings are thresholded to two levels with a threshold of 1.5. As discussed at the beginning of Sec. 2.4, this will indicate where the subject did and did not see differences in the image pair. Then optimal thresholds for the mean IFA array and maximum IFA array are selected. The optimal threshold is selected so that the number of coincidences between the psychophysical rating array and IFA array is maximized. By doing this the effect of subject's bias is likely to be minimized. In the multi-level case, the original psychophysical rating array is used to find four optimal thresholds for each IFA array. We obtain

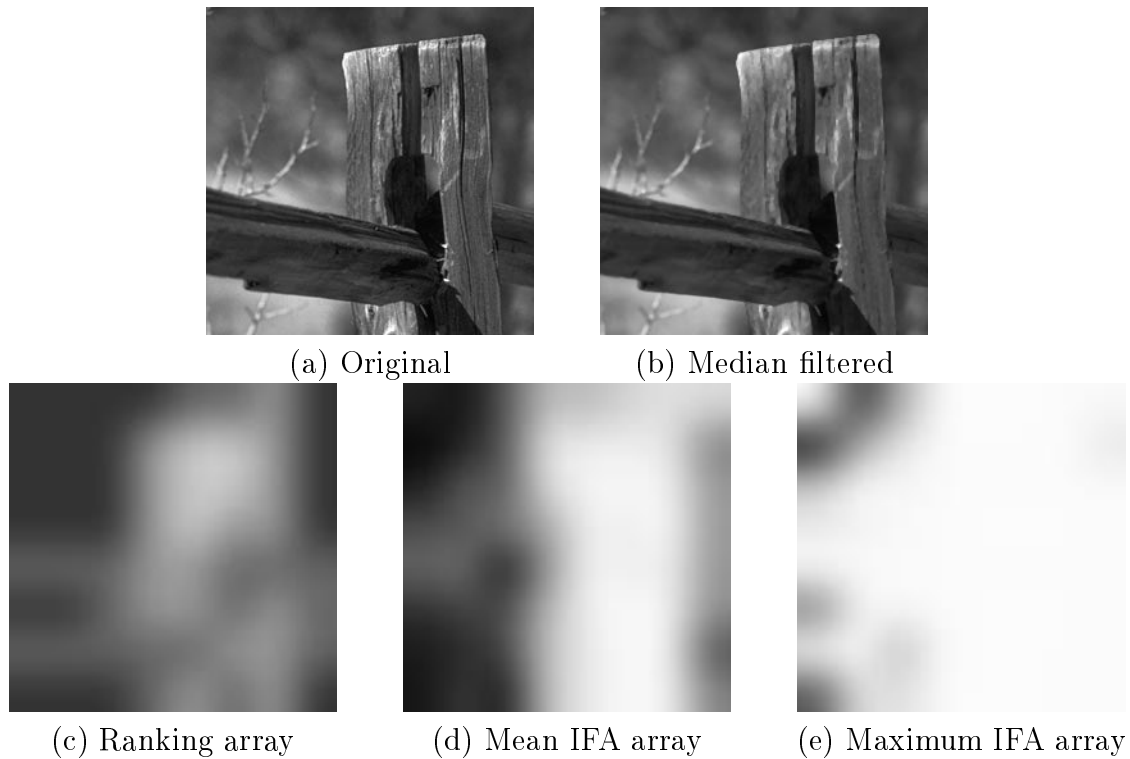


Figure 4. A visual comparison of psychophysical rating array, mean IFA array, and maximum IFA array for an image distorted by median filtering.

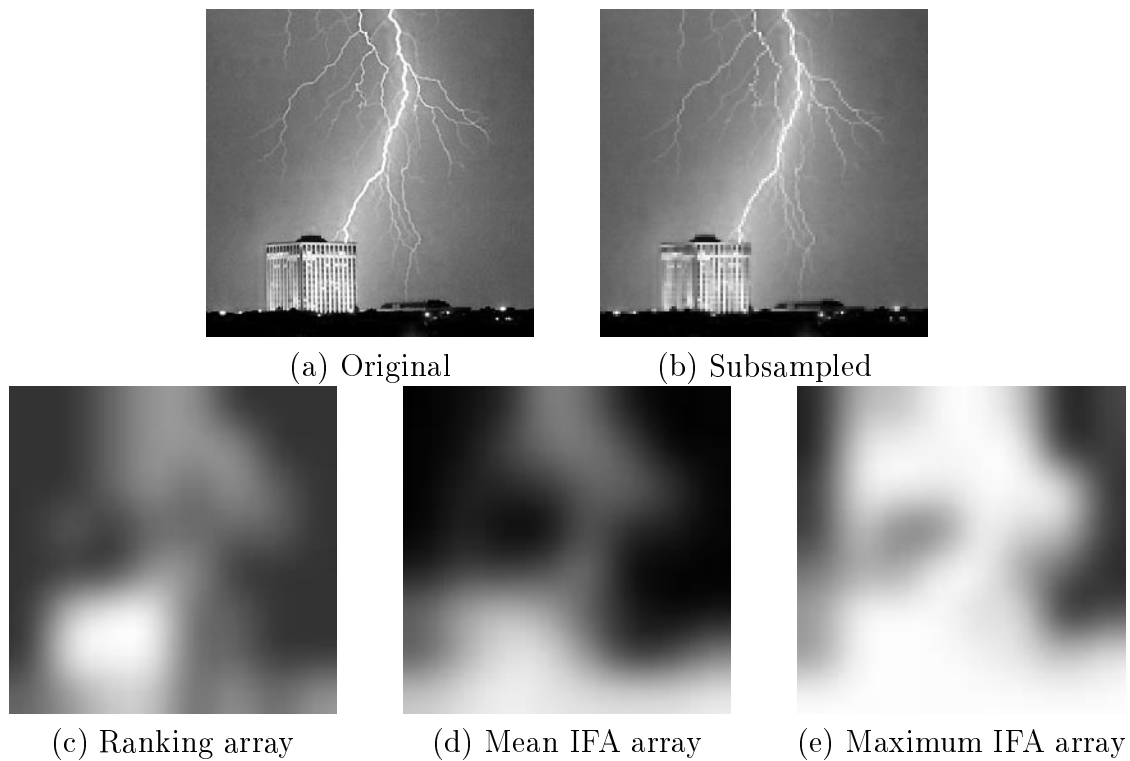


Figure 5. A visual comparison of psychophysical rating array, mean IFA array, and maximum IFA array for an image distorted by subsampling.

Image Name	Distortion Type	Bi-Level		Multi-Level	
		mean IFA array	max IFA array	mean IFA array	max IFA array
falls	sharpened	55.02	25.63	79.51	35.48
lightning	subsampling	33.27	25.60	88.88	56.18
parrot	crystallized	18.13	18.99	54.87	61.14
parrot	luminance shifted	26.56	35.62	85.59	95.80
parrot	median filtered	32.38	43.16	63.21	53.67
parrot	sharpened	23.04	25.09	63.50	36.65
pattern	subsampling	29.87	14.45	76.68	34.86
pattern	blurred	64.00	12.80	129.99	64.00
post	median filtered	18.32	7.61	36.19	17.36
shell	subsampling	31.84	19.96	36.39	25.55
Average		33.24	22.89	71.48	48.07

Table 1. Pearson independence test χ^2 values for the image pairs tested.

the thresholds sequentially in the same way as the bi-level case.

2.6. Results

Image fidelity measures are designed to match human perception. Therefore, we would expect the thresholded arrays from effective image fidelity measures to be highly consistent with the arrays obtained via psychophysical experiment. For each image pair, we perform a Pearson independence test⁹ on the psychophysical rating array and each IFA array. Table 1 contains the Pearson test χ^2 values for both the bi-level and multi-level approaches for all ten image pairs.

If the IFA prediction were random, the ratings in the IFA array would be independent of the psychophysical ratings. In this case, on the average, only one in ten χ^2 values would be above a value of 2.71 for the bi-level case and 23.54 for the multi-level case. Since all but one of the χ^2 values are above these critical values, we conclude that the IFA is producing results that are related to perceptual judgments.

2.7. Discussion

In order to test sensitivity of the method to changes in the model's quality, we chose to generate results for an impaired version of the IFA. The IFA was impaired by removing visual channels associated with specific orientations. Figure 6 shows that the removal of specific visual channels in the impaired IFA prevents the model from "seeing" image distortions along the North-East – South-West axis. The distortion maps produced by the impaired IFA were processed by the method described in the previous sections in order to obtain χ^2 values for the ten images pairs. It is expected that the χ^2 values for the impaired model would be lower than the χ^2 values for the unimpaired model. A comparison of the average of the χ^2 values for the IFA and impaired IFA models reveals that the IFA produces higher average χ^2 values (bi-level mean IFA array: 33.24 and multi-level mean IFA array: 71.48) than the impaired IFA (bi-level mean impaired IFA array: 31.17, multi-level mean impaired IFA array: 57.08). As expected, the proposed metric correctly identifies that the results of the IFA more closely match human perception than the results of the impaired IFA.

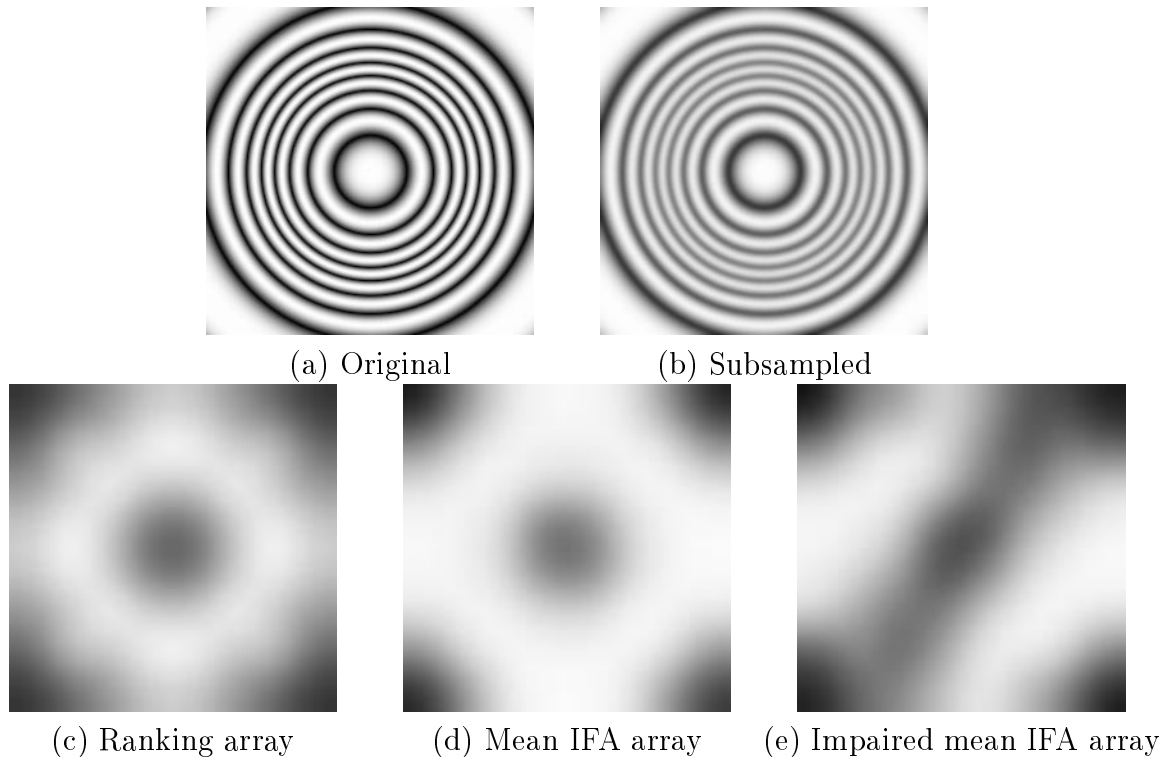


Figure 6. A visual comparison of psychophysical rating array, mean IFA array, and impaired mean IFA array for an image distorted by subsampling.

ACKNOWLEDGMENT

This work was supported by the Hewlett-Packard Company.

REFERENCES

1. S. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, ed., pp. 179 – 205, MIT Press, Cambridge, MA, 1993.
2. J. Lubin, "The use of psychophysical data and models," in *Digital Images and Human Vision*, A. B. Watson, ed., pp. 171 – 178, MIT Press, Cambridge, MA, 1993.
3. D. J. Heeger and P. C. Teo, "A model of perceptual image fidelity," in *Proc. of IEEE Int'l Conf. on Image Proc.*, pp. 343 – 345, (Washington, D.C., USA), Oct. 23 – 26 1995.
4. C. C. Taylor, Z. Pizlo, J. P. Allebach, and C. A. Bouman, "Image quality assessment with a Gabor pyramid model of the human visual system," in *Human Vision and Electronic Imaging*, vol. SPIE 3016, pp. 58 – 69, (San Jose, CA, USA), Feb. 8 – 14 1997.
5. S. Daly, "Quantitative performance assessment of an algorithm for the determination of image fidelity," in *SID 93 Digest*, pp. 317 – 320, 1993.
6. J. Lubin, "A visual discrimination model for imaging systems design and evaluation," in *Vision Models for Target Detection and Recognition*, E. Peli, ed., pp. 245 – 283, World Scientific, Singapore, 1995.
7. O. M. Blackwell and H. R. Blackwell, "Visual performance data for 156 normal observers of various ages," *J. Illum. Engr.* **61**, pp. 3 – 13, 1971.

8. X. Zhang, E. Setiawan, and B. Wandell, "Image distortion maps," in *Proc. of the 5th Color Imaging Conference: Color Science, Systems, and Applications*, pp. 120 – 125, (Scottsdale, AZ, USA), 17-20 November 1997.
9. R. E. Kirk, *Statistics, An Introduction*, Holt, Rinehart, and Winston, Inc., 6277 Sea Harbor Drive, Orlando, FL 32887, 3 ed., 1989.