Perceptual Assessment of Demosaicing Algorithm Performance

PHILIPPE LONGÈRE, XUEMEI ZHANG, PETER B. DELAHUNT, AND DAVID H. BRAINARD

Demosaicing is an important part of the image-processing chain for many digital color cameras. The demosaicing operation converts a raw image acquired with a single sensor array, overlaid with a color filter array, into a full-color image. In this paper, we report the results of two perceptual experiments that compare the perceptual quality of the output of different demosaicing algorithms. In the first experiment, we found that a Bayesian demosaicing algorithm produced the most preferred images. Detailed examination of the data, however, indicated that the good performance of this algorithm was at least in part due to the fact that it sharpened the images while it demosaiced them. In a second experiment, we silenced image sharpness as a factor by applying a sharpening algorithm to the output of each demosaicing algorithm. The optimal amount of sharpening to be applied to each image was chosen using the results of a preliminary experiment. Once sharpness was equated in this way, an algorithm developed by Freeman based on bilinear interpolation combined with median filtering, gave the best results. An analysis of our data suggests that our perceptual results cannot be easily predicted using an image metric.

Keywords*—*Demosaicing, digital camera, image quality.

I. INTRODUCTION

In most consumer digital cameras, the *full-color image* provided to the end user is specified by the responses of distinct red-green-blue (RGB) sensor classes at each pixel. Typically, however, the full-color image is not acquired directly. Rather, the camera captures the image using a single sensor array on which RGB color filters have been superposed. The use of interleaved color filters creates three distinct sensor classes, but at each individual pixel, the response of only one class is available. This is called a mosaiced design and a typical [1] filter arrangement is shown in Fig. 1. When a mosaiced design is used, the captured *raw image* must be processed to produce a full-color image. This processing is called demosaicing and consists of estimating at each pixel the responses of the sensor classes that are not directly available at that pixel.

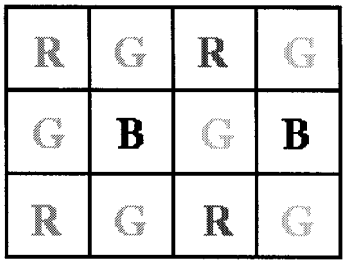

Fig. 1. Typical mosaic pattern. Each pixel in the overall sensor array is overlaid with a single color filter. Separate filter classes are labeled as R, G, and B to indicate red, green, and blue, respectively.

The demosaicing operation is usually implemented by the camera electronics and in this sense the mosaiced design and demosaicing operation are transparent to the end user. Nonetheless, it is well known that the use of a mosaiced design can lead to aliasing artifacts in the final demosaiced image [2]. Often, these artifacts manifest themselves as chromatic mottle or splotches in the vicinity of high spatial frequency luminance variation in the image (see Fig. 2). In general, the magnitude of the mosaicing artifacts depends on the image content, the optics of the camera, the pixel density of the sensor array, the mosaicing pattern, and the demosaicing algorithm applied.

A number of different demosaicing algorithms are available in the literature [3]–[9]. As shown in Fig. 2, use of different algorithms can lead to different demosaicing artifacts. In general, however, little is known about the relative efficacy of different demosaicing algorithms. This is particularly true if one is interested in the perceptual quality of the images provided to the end user. In this paper, we report the results of perceptual experiments designed to compare the efficacy of different demosaicing algorithms. For these experiments, we started with full-color images, mosaiced them by deleting two of the three RGB values at each pixel, and then demosaiced them with different algorithms.

Manuscript received March 12, 2001; revised July 16, 2001. This work was supported in part by the Hewlett-Packard Corporation and in part by Agilent Technologies, Inc.

P. Longère was with the Psychology Department, University of California, Santa Barbara, CA 93106 USA. He is now with m-Pixel, Sophia Antipolis, 06560 Valbonne, France (e-mail: phlongere@yahoo.fr).

X. Zhang is with the Agilent Technologies Laboratories, Palo Alto, CA 94303 USA. (e-mail: xmei@labs.agilent.com).

P. B. Delahunt was with the Psychology Department, University of California, Santa Barbara, CA 93106 USA. He is now with the University of California Davis Medical Center, Sacramento, CA 95817 USA (e-mail: pbdelahunt@ucdavis.edu).

D. H. Brainard was with the Psychology Department, University of California, Santa Barbara, CA 93106 USA. He is now with the Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: brainard@psych.upenn.edu).

Publisher Item Identifier S 0018-9219(02)00808-3.

Fig. 2. Example results from different demosaicing methods. (a) Full-color original image is compared to the results of applying (b) and (c) two different demosaicing algorithms to a mosaiced version of the original. Full-color image itself was obtained by subsampling a high-spatial resolution mosaiced image acquired with a Kodak DCS200 (see Section III-A). Mosaicing artifacts can be observed in the two demosaiced images, but the output of the two algorithms differs in detail.

The rest of this paper is organized as follows. In Section II, we describe the demosaicing algorithms whose performance we studied. Then we describe the experimental methods we used to assess image quality and the results of two experiments. In the first experiment, we evaluated the performance of four demosaicing algorithms. These were used in conjunction with a simple color balancing algorithm that served to render the RGB camera images for display on a color monitor. One striking result of the first experiment was that demosaiced images sometimes appeared better than the full-color images from which they were derived. Visual examination of the images suggested that this effect arose because some of the demosaicing algorithms not only estimated missing sensor values, but also sharpened the image. In the second experiment, we equated the sharpness of the experimental images by adding a sharpening operation to the image-processing pipeline. This operation served to separate demosaicing per se from other effects of the demosaicing algorithms. After we present the results of the second experiment, we conclude with some general comments.

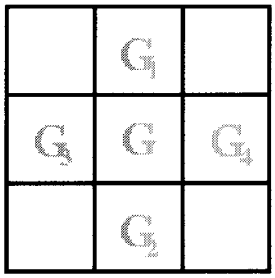

Fig. 3. Pixels used for bilinear interpolation of green image plane. Missing green value at the center pixel (labeled G) is obtained by taking the mean of the known green value at the neighboring locations where there is a green sensor (labeled G_1, G_2, G_3, G_4).

II. Demosaicing Algorithms

The demosaicing algorithms used in most commercial cameras are proprietary. For the experiments presented here, we studied four demosaicing algorithms that were available to us. These are described below.

A. Bilinear Demosaicing

Bilinear interpolation may be used as a demosaicing algorithm if each color plane is treated independently. It is perhaps the simplest demosaicing algorithm and, thus, serves as a useful baseline for comparison. Fig. 3 illustrates bilinear interpolation for the G sensor class of the mosaic pattern shown in Fig. 1. Each pixel without a G sensor is surrounded by four pixels with G sensors as shown in Fig. 3. Thus, the missing value may be obtained from the surrounding G sensor responses as: $G = (G_1 + G_2 + G_3 + G_4)/4$. For the R and B sensor classes, missing values are similarly estimated as linear combinations of available R and B sensor responses, respectively. Bilinear demosaicing has the feature that each estimated value depends only on responses from the same sensor class. It would provide a near-optimal solution if the R, G, and B responses were statistically independent, as then responses from one sensor class would provide no predictive information about responses in another. For natural im-

ages and typical choices of R, G, and B sensor spectral sensitivities, however, such statistical independence of responses across sensor classes does not hold [2], [10]–[12]. The other demosaicing algorithms we considered attempt to take advantage of the correlations between responses in different sensor classes.

B. Freeman Demosaicing

In the mosaic pattern shown in Fig. 1, the G sensor class is sampled at a higher rate than the R and B sensor classes. Thus, one might expect that the result of bilinear interpolation would be more accurate for the G sensor class than for the R and B sensor classes. Freeman's [6] algorithm is designed to take advantage of this intuition by using the interpolated G image to modify the interpolated R and B images. The algorithm begins with bilinear interpolation applied separately to each sensor class, as described above. The G image plane resulting from bilinear interpolation is then used directly as the output G image plane. For the R and B image planes, however, the result of bilinear interpolation is modified before output. First two difference image planes R-G and B-G are created. These may be thought of as representations of the chromatic content of the image. Since mosaicing artifacts are generally manifest as small chromatic splotches, median filtering the R-G and B-G planes will tend to eliminate them. In the Freeman algorithm, such median filtering is applied. Finally, R and B output planes are created by adding the G plane to the median filtered R-G and B-G planes. In the final output image, the R and B values estimated from the median filtered R-G and B-G planes are used only at pixels where there is no R or B sensor value directly available. In our implementation of the Freeman algorithm, we used a 5 $\times$ 5 median filter.

As reported below, the performance of the Freeman algorithm in Experiment 1 was very bad. We determined that the cause of this was that the R, G, and B sensor classes had very different mean responses in our images. The patented Freeman algorithm is sensitive to this differential scaling. Scaling the R, G, and B sensor planes to have the same mean before applying the Freeman algorithm and undoing this scaling afterwards produced much better results and we used this modification in Experiment 2. We refer to the two versions of the Freeman algorithm as Freeman 1 and Freeman 2, respectively. A visual comparison of their performance is shown in Fig. 4.

C. Bayesian Demosaicing

The Bayesian demosaicing method we used is described in more detail elsewhere [8], [9], [13]. The basic idea is to use the raw image data together with prior information about the spatial and chromatic structure of natural images to obtain the demosaiced full-color image. In the Bayesian approach, scene parameters (in our case, the full-color image) are estimated from observed data (in our case the raw image) together with a prior distribution over the scene parameters. To implement a Bayesian algorithm, one must specify both the prior and a likelihood function which describes the relation between the scene parameters and the observed data.

(a) Freeman 1

(b) Freeman 2

Fig. 4. Comparison of (a) Freeman 1 and (b) Freeman 2 demosaicing algorithms. The difference between the two versions is that in the Freeman 2 algorithm the output of each sensor class is normalized before demosaicing and rescaled after demosaicing. Demosaicing artifacts are considerably more salient in the output of the Freeman 1 algorithm.

For notational convenience, we denote the scene parameters by the vector x. The data available for estimation are the responses of the mosaiced sensor array, which we denote by the vector y. The likelihood is specified by the probability distribution $p(y|x)$: this gives the probability that image data y will be recorded given that the full-color image was x. The prior is specified by a distribution $p(x)$: this gives the probability that any particular full-color image x was present when the data were acquired. Given the likelihood and the prior, Bayes' rule may be used to compute a posterior distribution $p(x|y) = Cp(y|x)p(x)$, where C is a normalizing constant. The posterior yields how likely it is that the full-color image was x given observed data y. An estimate for the full-color image may then be obtained from the posterior distribution, e.g., by taking its mean.

In Bayesian image processing, specification of the likelihood function is generally straightforward. In our case, we incorporated optical blur in the form of a point spread function, a small amount of additive sensor noise, and the mosaic pattern itself into the computation of the likelihood. The point spread function was taken as a circularly symmetric Gaussian with a standard deviation of one pixel.

Specification of the prior is the second important component in developing a Bayesian algorithm. Given our current limited understanding of the statistical properties of natural scenes, specifying a prior involves as much art as it does science. We assumed that the full-color images were separable in space and color, so that: 1) the color statistics at each pixel were governed by the same distribution and 2) the spatial statistics in each color plane were the same. The color statistics were taken to be trivariate Gaussian. The spatial statistics were also taken to be multivariate Gaussian, with covariance matrix based on the assumptions that: 1) the spatial properties of the image are separable in the row and column dimensions and 2) that in each dimension, the spatial properties are characterized by a first-order Gauss–Markov process with the same space parameter for the row and column dimensions. Given these assumptions, the spatial statistics are specified by the correlation between nearest neighbor pixels within each image plane. This method of specifying full-color image priors is described in detail in a separate report [13].

We studied two versions of the Bayesian algorithm. They differed primarily in how the parameters of the prior distribution were obtained. In the first version (Bayesian 1), we set the prior parameters using a bootstrapping method.[1] First, we used bilinear demosaicing to produce a full-color version of the image. We then analyzed the result of bilinear interpolation to obtain the mean and covariance matrix of the RGB responses. These parameters determined the color portion of the prior. In addition, we obtained the average correlation between neighboring pixels in each image plane. Given the Gauss–Markov assumption, this single parameter determined the spatial portion of the prior. In the second version (Bayesian 2), the color priors were set in the same way as for the first. The spatial parameter, however, was simply set to a value of 0.70, considerably lower than the value found in any of the images we examined. Specifying a lower correlation has the effect of increasing the prior probability of high spatial frequencies in the image.[2] For both the Bayesian 1 and Bayesian 2 algorithms, the estimates at each pixel were obtained from the mean of the posterior obtained when the Bayesian analysis was applied to a 5×5 image region surrounding that pixel.

It is worth noting the Bayesian algorithms could modify sensor values for sensor classes and locations where a directly measured value was in fact available in the raw image.

[1]We would like to thank D. Keren for suggesting the bootstrapping approach.

[2]In addition, the Bayesian 2 algorithm assumed a circular symmetric rather than an $x-y$ separable Gauss–Markov model, but this change had only a minor effect on the algorithm's performance.

Fig. 5. Experimental setup. On each trial, the subject viewed two images of the same scene. Images had been processed differently. Subject indicated which image appeared the most appealing by clicking the mouse on the box below one or the other image.

III. Experiment 1

A. Methods

There are a number of perceptual criteria that could be used to evaluate the performance of demosaicing algorithms. For example, one could ask which algorithm produced images that were most perceptually similar to the true full-color image. Our interest was in evaluating algorithms for use in consumer cameras, however, and we thought that the appropriate perceptual question to ask was which algorithm produced images that were most pleasing to the subjects. The following paragraphs describe our experimental methods.

Observers viewed images presented on a 21-in monitor (HP P1100) in a room illuminated by fluorescent ceiling lights. These viewing conditions were chosen because they seemed typical of the conditions under which many consumers would view digital photographs. On each trial of the experiment, subjects viewed two images of the same scene (see Fig. 5). The two images had been processed differently and the subject was asked to indicate which image was the most appealing in the sense that it was the one he or she would prefer to put in a photo album. Subjects were not allowed to respond until after they had viewed the images for at least two seconds.

We used five different images of each scene. We refer to one of the images as the *nonmosaiced* image. This image was obtained by subsampling a high spatial resolution mosaiced image. The high-resolution images were acquired in Santa Barbara, CA, and Palo Alto, CA, under various daylight conditions using a Kodak DCS200 camera and a 50-mm lens. The native resolution of the DCS200 is 1524×1012 pixels. The interface software for the camera allows access to the data in raw form: no demosaicing, no color correction, no gamma correction, and no tone mapping have been applied to this raw data (see [14]). Before subsampling, a proprietary algorithm was then applied to the raw image data to eliminate

artifacts that were present in some of the images because of saturated responses at individual pixels. This procedure only affected the values of saturated pixels. Each color plane of the resulting high-resolution raw image was then separately downsampled by a factor of two using bilinear interpolation. By reducing the resolution, we ensure that color sensors from each class sample the image at a high enough rate to minimize mosaicing artifacts in the subsampled image. After subsampling, some of the images were cropped so that they could be seen side by side at the native 1280 $\times$ 1024 pixel resolution of our graphics display hardware.

The other four images of each scene were derived from the nonmosaiced image. The procedure was to first mosaic the nonmosaiced image according to the pattern shown in Fig. 1 and then to apply one of the demosaicing algorithms described in the previous section. For Experiment 1, we created *Bilinear*, *Freeman 1*, *Bayesian 1*, and *Bayesian 2* images for each scene by applying the corresponding algorithm.

Before use in the experiment, each image was rendered for display on our monitor using standard methods for color balancing and gamma correction [2]. Images of ten different scenes were used. All of the images used in Experiment 1 are available for download on the world wide web.[3] Note that in our image-processing pipeline, demosaicing was performed first followed by color balancing and gamma correction. The images on the web site are those used in the experiment, not those input to the demosaicing algorithms.

During an experimental session, each of the images of a particular scene was presented once in conjunction with each of the other images of that scene. Thus, there were a total of 100 trials per session (ten scenes and ten possible comparisons per scene). The order of presentation of the 100 possible trials was randomized in each session. On each trial, which image was presented on which side of the display was also chosen randomly.

Nine subjects participated in Experiment 1. Each subject participated in one experimental session. All subjects had color-normal vision as tested with the Ishihara plates [15] and all reported that they had normal or corrected to normal visual acuity. The subjects were paid volunteers recruited by advertisement on the University of California, Santa Barbara, campus. All were undergraduate students. Although the subjects were informed that the purpose of the experiment was to improve our understanding of how image processing affects image quality, all were naive as to the experimental design and the nature of the image processing under study. Indeed, no instructions were given about the types of artifacts likely to be visible in the images, so that subjects would not artificially focus on any particular aspect.

Subjects viewed the screen from 60 cm. At this distance, the full 1280 $\times$ 1024 pixel display subtended 36 $\times$ 29° of visual angle. The size of the individual images varied from image to image. Across all the experiments, the smallest image linear dimension was 12° and the largest was 22°.

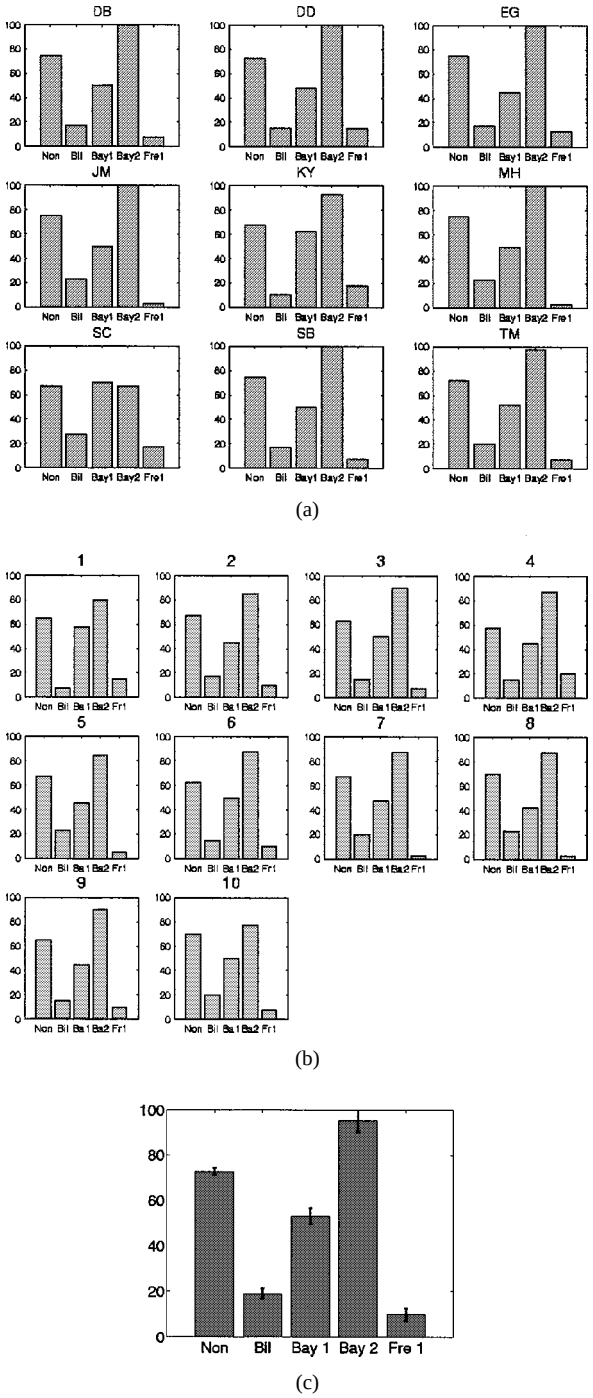

Fig. 6. Experiment 1 results. (a) Preference by subject, averaged over images. Preference score is the percentage of times a given image type was chosen as preferred. Along the abscissa, the image types are ordered as follows: nonmosaiced, bilinear, Bayesian 1, Bayesian 2, and Freeman 1. (b) Preference by image, averaged over subjects. Image numbers here and elsewhere in the paper refer to those used to identify individual scenes are available.[4] (c) Preference scores averaged over both images and subjects. Error bars show $+/-1$ SEM.

B. Results

For each subject and scene, the results of the experiment may be expressed as the percentage of times that the result of each method was chosen on trials where it was presented.

[3]http://color.psych.upenn.edu/depreference

[4]http://color.psych.upenn.edu/depreference

Thus, each method can be assigned a score between 0% and 100%. The higher the percentage, the more preferred the method. Fig. 6 shows the preference scores obtained in Experiment 1.

Fig. 6(a) shows data for each subject, averaged over all the scenes. There is good consistency across subjects. For seven of the nine subjects, the rank ordering of the image types was the same: Bayesian 2 (most preferred), Nonmosaiced, Bayesian 1, Bilinear, and Freeman 1 (least preferred). For subject SC, Bayesian 1 was the most preferred algorithm. For subject KY, Freeman 1 was preferred to Bilinear.

Fig. 6(b) shows data for each scene, averaged over all the subjects. There is also good consistency across scenes. This view of the data tells a very similar story. The Bayesian 2 is the most preferred image type for all scenes, followed by Nonmosaiced and Bayesian 1. The Bilinear and Freeman 1 algorithms performed poorly. Which one produced the least favored results varied from scene to scene.

Fig. 6(c) shows the overall mean data along with the standard errors of the mean. The overall ordering computed in this way matches that shown by most individual subjects: Bayesian 2 (most preferred), Nonmosaiced, Bayesian 1, Bilinear, and Freeman 1 (least preferred). Although the Freeman 1 algorithm performed worse on average than the Bilinear algorithm, the individual scene analysis above suggests that this ordering will vary with the choice of scenes studied in the experiment.

C. Discussion

The results of Experiment 1 show a great deal of consistency. This tells us that there is sufficient agreement across both individual subjects and across scenes for meaningful statements to be made about algorithm performance. If it had been the case that the rank ordering of the different algorithms varied greatly across either subjects or scenes, then we would have learned that it is not possible to make general statements about which algorithm is best. Such a situation would greatly complicate optimization of camera design because it would be necessary to match the image processing to the preferences of individual consumers and to vary the processing depending on scene content. Another way to say this is that it is possible to obtain meaningful image preference data with a modest experimental investment. It does remain possible that more individual variability would have been observed if we had used a more heterogeneous population of subjects (e.g., wider range of cultural backgrounds or ages).

Two aspects of the experimental results surprised us. First, the Bayesian 2 algorithm produced images that were more pleasing than the nonmosaiced images from which they were derived. Second, the Freeman 1 algorithm performed very badly, doing no better than simple bilinear demosaicing. This surprised us since the Freeman 1 algorithm was designed explicitly to remove artifacts present after bilinear demosaicing.

As noted above, the failure of the Freeman 1 algorithm to perform well led us to look in more detail at our implementation. Although we had correctly implemented the algorithm as described in the patent [6], that implementation did not perform well with our particular camera. It was the poor performance of the Freeman 1 algorithm that motivated us to develop the Freeman 2 algorithm. This version handles a more diverse set of camera sensor designs than the original version and we used it in Experiment 2 below.

We also tried to understand how it could be that a mosaiced and then demosaiced image could be preferred to the original nonmosaiced image. After obtaining the results of Experiment 1, we examined the individual images in some detail. Although the Bayesian 2 images contained mosaicing artifacts not present in the nonmosaiced images, the Bayesian 2 images also appeared to be sharper than their nonmosaiced counterparts. Image sharpness (or blurriness) is known to be an important factor influencing image quality judgments. [16]–[18]

Recall that the Bayesian algorithms try to estimate the incident full-color image and that the likelihood function we used contained information about the optical point spread of the camera. Because of this the Bayesian processing not only demosaics the image, it also sharpens it. In the Bayesian 2 algorithm, the correlation parameter of the spatial component of the prior was set to a low level, indicating that high spatial frequencies were likely to be present in the incident image. Such a choice of prior has the effect of encouraging the algorithm to sharpen the image. Apparently, the good performance of the Bayesian 2 algorithm was driven by its sharpening properties as well as its demosaicing performance. Note that the sharpening accomplished by the Bayesian algorithm is deeply embedded in its design, so that it is not obvious that one can perform optimal Bayesian demosaicing without also allowing the algorithm to sharpen the image. Although we could manipulate parameters such as the specified optical point spread function (see below) to reduce the amount of sharpening performed by the algorithm, this would also affect the algorithm's demosaicing performance.

The results of Experiment 1 led us to conclude that a fair comparison of demosaicing methods ought to be one that silenced the role of image sharpness in the preference judgments. Thus, we conducted Experiment 2.

IV. Experiment 2

A. Introduction

The purpose of Experiment 2 was to compare demosaicing methods in a manner that eliminated the role of image sharpness. Since we did not know how to prevent the Bayesian algorithms from sharpening the images (see Section III), we decided to postprocess all the images with a sharpening algorithm so that sharpness was equated. This procedure also compensates for any blurring or sharpening implicit in the other demosaicing algorithms. Some degree of such blurring or sharpening is to be expected, since all the algorithms combine information across neighboring pixels to accomplish demosaicing. Our hope was that by sharpening all the images, we would better isolate the demosaicing performance of the individual algorithms.

A number of different algorithms are available for sharpening images. Most of these are based on spatial or spectral filtering of the images to enhance high frequencies. It is beyond the scope of this paper to provide a survey of sharpening methods. Information on such methods is available elsewhere [17], [19]. What we required was a method that allowed us to parametrically vary the degree of sharpening performed by the algorithm. Since we had already discovered that the Bayesian approach provided an effective means of sharpening images, we modified our Bayesian demosaicing algorithm so that it could also be applied to full-color images. To vary the amount of sharpening applied by the algorithm, we varied the size of the optical point spread function used to compute the likelihood function $p(y|x)$ fed to the Bayesian calculation. Specifying a large point spread function has the effect of telling the algorithm that more high-frequency information has been removed from the raw image data and, thus, increases the extent to which the algorithm attempts to restore such information in the estimated full-color image.

B. Preliminary Experiment

The relationship between sharpness and image quality is not trivial. If one starts with a blurry image and applies a moderate amount of image sharpening, image quality typically improves. As the degree of sharpening is increased further, however, image quality will deteriorate (see Fig. 7). The optimal amount of sharpening to apply varies from image to image and we are unaware of methods that automatically determine the optimal amount. Since we wanted to compare images when each had been optimally sharpened, we conducted a preliminary experiment to determine this optimal amount for each of our experimental images. Fig. 8 shows the results for three scenes. Note that the optimal amount of sharpening varies both across scenes and across image types. It makes sense that less sharpening would be required for the output of the Bayesian 2 method, since, as noted above, this method has the intrinsic property of substantially increasing image sharpness. The variation across algorithms may arise because each effectively applies a different level of intrinsic blurring or sharpening. Although there was some variability between subjects in the preferred amount of sharpness, we took the average across subjects to determine the optimal level of sharpening for each individual image to be used in Experiment 2. These optimal levels should be regarded as specific to the resolution of our image set and display device. In addition, note that the optimally sharpened version of even the nonmosaiced image is not artifact free. Rather, it represents the best tradeoff between the benefits of sharpness and the costs of other artifacts introduced through the sharpening process.

In the preliminary experiment, five subjects adjusted image sharpness so that the adjusted image appeared most pleasing. For each adjustment, subjects could choose the degree of sharpness by selecting one of eight precomputed versions of the same image. The precomputed images were shown one at a time on the same equipment used in the main experiments. The precomputed sharpness levels were chosen to span a range of sharpening from no sharpening to obviously oversharpened (as judged by the first author).

(a) Original Image

(b) Optimal Sharpening Setting

(c) Maximum Sharpening Level

Fig. 7. Effect of varying image sharpness. (a) Portion of an unsharpened nonmosaiced image. (b) Same image area optimally sharpened as determined by our preliminary experiment. (c) Oversharpened version of the same image area. Oversharpened version corresponds to the largest specified point spread function used. Note that this version does not have the appearance typical of oversharpening with other algorithms (e.g., unsharp masking). Rather, the oversharpened image appears somewhat blurred.

Sharpening was applied after demosaicing and before color balancing and gamma correction.

We ran the preliminary experiment for each image type (Nonmosaiced, Bilinear, Bayesian 1, Bayesian 2, and Freeman 2) of each scene to be used in Experiment 2. The five subjects in the preliminary experiment were the first

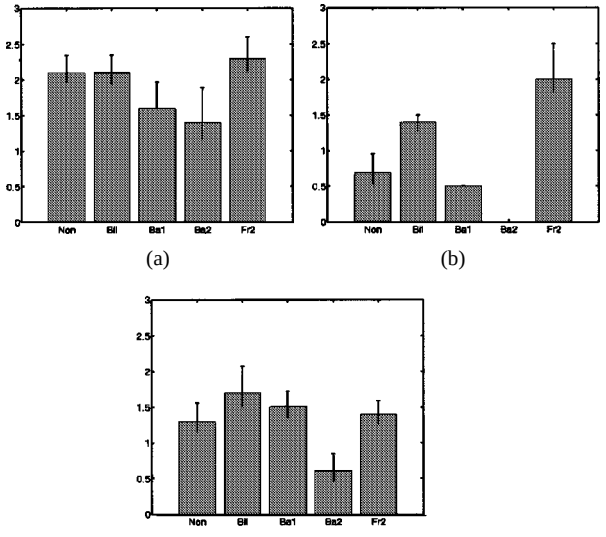

Fig. 8. Results of preliminary experiment. (a) Scene 1. (b) Scene 4. (c) Scene 11. Each panel shows the results of the preliminary experiment for one scene. Each bar shows the optimal sharpening level for a single image type. Along the abscissa the image types are ordered as follows: Nonmosaiced, Bilinear, Bayesian 1, Bayesian 2 and Freeman 2. The ordinate provides the standard deviation of the circularly symmetric Gaussian point spread function expressed in pixels. The error bars show $+/-1$ SEM.

author, the last author, and three members of the laboratory. These subjects were not particularly naive, but for this purpose, we did not feel this was a disadvantage. The sharpness judgments were made for all individual images in a single experimental session. Within the session, individual images (different scenes and image types) were presented once each in random order.

C. Methods

The methods for Experiment 2 were identical to those of Experiment 1. As described above, the images were sharpened after demosaicing and before they were rendered for monitor display. Nine rather than ten scenes were used. We reduced the number of scenes to shorten the experimental session so that subjects could finish in a one hour time block. Eight of the scenes were also used in Experiment 1. The nonmosaiced version of the ninth scene was acquired with a Nikon D1 digital camera with a 28–70-mm zoom lens. The native resolution of the D1 is 1324×2012 pixels. The D1 also allows access to the image data in raw form, although it was necessary to undo a white balancing operation that had been applied by the camera. The images shown to subjects in Experiment 2 are also available at the web site.

In Experiment 2, the Freeman 2 algorithm was substituted for the Freeman 1 algorithm.

Ten additional naive subjects participated in Experiment 2. Again, these were all undergraduate students. Data from one subject were excluded because a subsequent test with the Ishihara plates revealed that he was color deficient. The other nine subjects were color normal as assessed by the plates and had normal or corrected to normal visual acuity by self-report.

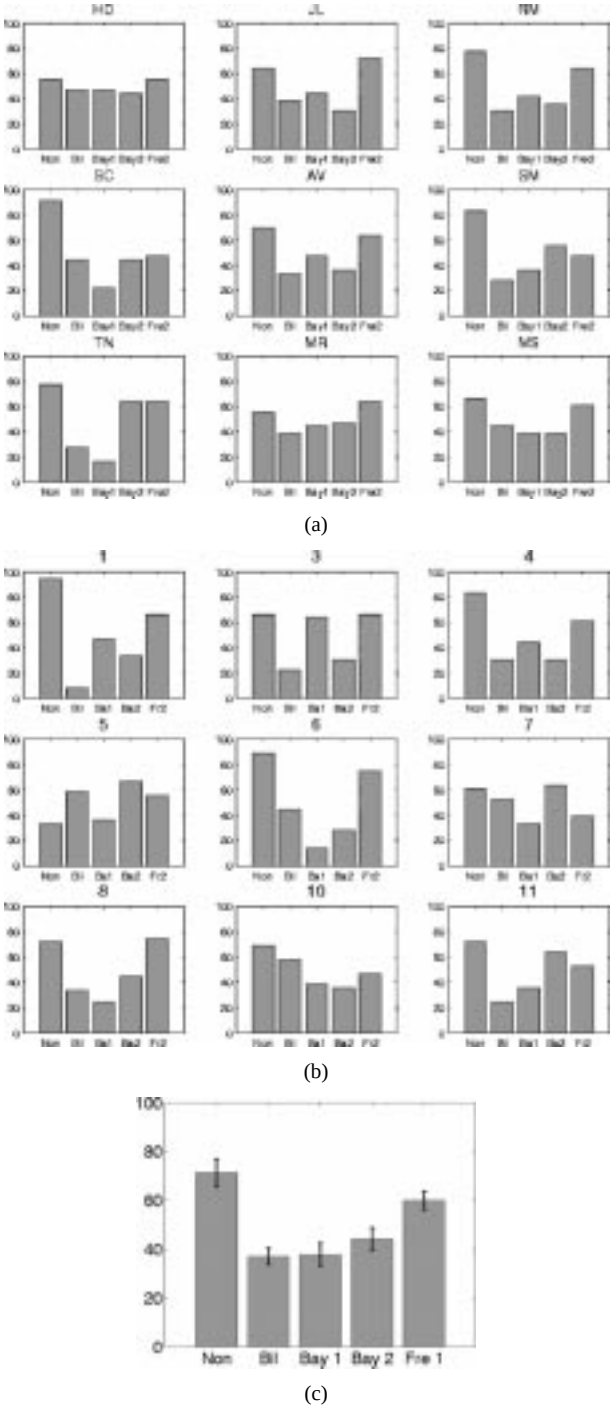

Fig. 9. Experiment 2 results. Same format as Fig. 6. (a) Preference by subject. (b) Preference by image. (c) Average preference.

D. Results

The results of Experiment 2 are presented on Fig. 9 in the same format as Fig. 6. A survey of the figure reveals more variation across subjects and images than was seen in Experiment 1. We believe this occurred because once sharpness had been equated, the differences between the algorithms' outputs were more subtle. This idea is supported by the fact that the range of quality scores across algorithms is compressed in Experiment 2 relative to Experiment 1, both for individual subject and individual image data.

Although there is less consistency in Experiment 2, there is still considerable regularity in the data. In general, the nonmosaiced image is preferred to all of the mosaiced–demosaiced versions. This statement is true for seven of nine individual subjects, five of nine individual images, and for the mean data. The fact that the nonmosaiced image looks good to subjects is reassuring and suggests that our sharpening operations worked as intended. Of the demosaicing algorithms, the Freeman 2 algorithm was the most preferred, followed by the Bayesian 2, Bayesian 1, and Bilinear algorithms, respectively. The differences between algorithms were small, however. Note that the standard errors of the means are large compared to the differences between Bilinear, Bayesian 1, and Bayesian 2 algorithms.

V. Discussion

We have described experiments that compare the perceptual quality of the output of various image-processing algorithms. In our experiments, we defined perceptual quality in terms of which image subjects found most pleasing in the sense that they would most like to put it in their photo album. Our main focus was to compare the efficacy of demosaicing algorithms. Here, we summarize the main conclusions from our experiments.

First, image preference experiments of the sort presented here are feasible and give meaningful results: when the differences between different images of the same scenes are large enough (as in Experiment 1), the results are very regular both across subjects and across images. Note that the image differences in Experiment 1 were within the range that one would routinely encounter when comparing algorithm performance—the images used in this experiment represented the output of reasonable algorithms. Indeed, the results of Experiment 1 helped us to improve several algorithms and led us to a second round of experimentation in which the differences between algorithms were narrowed.

Second, we learned in Experiment 1 that when comparing image-processing algorithms, it is important to be aware of the possibility that preference judgments are not based on the aspect of the image of interest to the experimenter. In Experiment 1, we intended to learn about the detrimental effect of mosaicing artifacts, but in the event subject judgments were driven by image sharpness. This was possible because sharpening was a collateral side effect of some of the demosaicing algorithms. We believe this point is likely to be of general relevance when comparing image-processing algorithms. Results of image preference experiments are probably most useful when used in conjunction with careful examination of the images themselves. Surprising results in the preference experiments may lead to insights for algorithm development. In our case, we learned that sharpened images with demosaicing artifacts (i.e., Bayesian 2 images) looked better than nonsharpened images with no artifacts (i.e., nonmosaiced images). This fact directed our attention to combining demosaicing with image sharpening.

Third, once image sharpness was equated, the differences between demosaicing algorithms were fairly small for our image set. Nonetheless, the Freeman 2 algorithm was the most effective of those we considered.

Table 1
S-CIELAB Distances Between NonMosaiced and Demosaiced Images

	Bay1	Bay2	Bili	Free
Experiment 1	4.0640	5.3218	2.5190	3.3573
Experiment 2	3.7409	4.4663	3.3301	2.7216

Distances Were Computed Using the Displayed Images and With Respect to the Monitor's White Point. Display MTF was not Measured or Taken Into Account in Computation of Distances.

One thing we did not do in our experiments was systematically vary the properties of the nonmosaiced images themselves. Nor did we vary the resolution of the display. Given our current rudimentary understanding both of demosaicing and of image preference, it is probably wise not to generalize our conclusion about which algorithm works best much beyond the particular cameras and display we studied. Camera optical quality, pixel density, sensor spectral sensitivities, mosaic pattern, pixel noise, and display resolution could all plausibly interact with the choice of demosaicing algorithm. In addition, the interaction between sharpening and demosaicing probably depends on the particulars of image acquisition. In particular, variation in the focus and depth of field of the nonmosaiced image could have a substantial effect on the results. For any particular choice of camera design parameters, however, our methodology can be used to evaluate algorithm performance for any scenes of interest. We have recently invested considerable effort in the development and validation of a digital camera simulator [2] so that such experimentation could be performed without having to build camera prototypes.

We close with a few remarks about the relation between our data and image metrics. Although we have found that it is feasible to conduct psychophysical studies on image preference and to use these as part of algorithm development, it would be desirable if it were possible to compute the perceptual quality of an image by comparing it to a baseline image using some sort of image comparison metric. For example, if it were the case that the most preferred images were the ones most similar (according to some metric) to the nonmosaiced version of the image, then algorithm evaluation would be simplified because it would not be necessary to run the perceptual experiments. The results of Experiment 1 clearly falsify this possibility, since the Bayesian 2 image is in fact preferred to the nonmosaiced image. Moreover, when we computed the average distance between the output of each of the algorithms and the nonmosaiced image using the S-CIELAB[20] image metric, the Bayesian 2 algorithm had the highest average distance, followed by the Bayesian 1, Freeman 1, and Bilinear algorithms in that order (see Table 1): the images most different from the nonmosaiced version were the most preferred. Thus, for the images and algorithms studied in Experiment 1, the use of the S-CIELAB metric (and probably any other current metric) in conjunction with a baseline nonmosaiced original image would provide a highly inaccurate prediction of actual image quality.

In Experiment 2, sharpness was removed as a salient factor and the nonmosaiced image was the most preferred, making it more plausible that an image metric could predict the results. Here, the differences between images were mainly the mosaicing artifacts. We evaluated whether the S-CIELAB color image metric (between the sharpened output of each algorithm and the sharpened nonmosaiced image) predicted the rank ordering of the algorithms. Again, the answer was no. Although the Freeman 2 algorithm did have the smallest average error, the Bayesian 2 algorithm, which was the next most preferred, had the largest. Perhaps this should not be surprising, given that the S-CIELAB metric is derived from threshold-level judgments of image difference, rather than the sort of image preference judgments used here. It is also possible that better prediction of preference would be obtained with current metrics if instead of comparing to the actual nonmosaiced image one compared to an hypothetical ideal (most preferred) baseline image of the scene.

ACKNOWLEDGMENT

The authors would like to thank Z. Baharav, J. Farell, W. Freeman, D. Keren, D. Sherman, and three anonymous reviewers for useful discussions or comments.

REFERENCES

[1] B. E. Bayer and Eastman Kodak Company, "Color Imaging Array," US Patent 3 971 065, 1975.
[2] P. Longère and D. H. Brainard, "Simulation of digital camera images from hyperspectral input," in Vision Models and Applications to Image and Video Processing, C. J. van den Branden Lambrecht, Ed. Norwell, MA: Kluwer, 2001.
[3] D. R. Cok and Eastman Kodak Company, "Signal processing method and apparatus for producing interpolated chrominance values in a sampled color image signal," US Patent 4 642 678, 1986.
[4] R. Kimmel, "Demosaicing: Image reconstruction from color CCD samples," *IEEE Trans. Image Processing*, vol. 8, pp. 1221–1228, Sept. 1999.
[5] B. Tao, I. Tastl, T. Cooper, M. Blasgen, and E. Edwards, "Demosaicing using human visual properties and wavelet interpolation filtering," in *Proc. IS&T/SID Seventh Color Imaging Conf.*, 1999, pp. 252–256.
[6] W. T. Freeman and Polaroid Corporation, "Method and apparatus for reconstructing missing color samples," US Patent 4 774 565, 1988.
[7] M. A. Wober, R. Soini, and Polaroid Corporation, "Method and apparatus for recovering image data through the use of a color test pattern," US Patent 5 475 769, 1995.
[8] D. H. Brainard. (1995) An ideal observer for appearance: Reconstruction from samples. Dept. Psychology, Univ. California, Santa Barbara, CA. [Online]. Available: http://color.psych.upenn.edu/brainard/papers/bayessampling.pdf
[9] D. H. Brainard and D. Sherman, "Reconstructing image from trichromatic samples: from basic research to practical applications," in *Proc. IS&T/SID 1995 Color Imaging Conf.*, 1995.
[10] G. Buchsbaum and A. Gottschalk, "Trichromacy, opponent colors coding and optimum color information transmission in the retina," *Proc. R. Soc. London*, vol. 220, no. 1248, pp. 89–113, 1983.
[11] D. L. Ruderman, T. W. Cronin, and C. C. Chiao, "Statistics of cone responses to natural images: implications for visual coding," *J. Opt. Soc. Amer. A*, vol. 15, no. 8, pp. 2036–2045, 1998.
[12] B. A. Wandell, *Foundations of Vision*. Sunderland, MA: Sinauer, 1995.
[13] D. H. Brainard, "Bayesian method for reconstructing color images from trichromatic samples," in *Proc. IS&T 47th Annual Meeting*, Rochester, NY, 1994, pp. 375–380.
[14] P. L. Vora, J. E. Farell, J. D. Tietz, and D. H. Brainard, "Image capture: simulation of sensor responses from hyperspectral images," *IEEE Trans. Image Processing*, vol. 10, pp. 307–316, Feb. 2001.
[15] *Ishihara's Design Charts for Color Deficiency*, Ishihara & Co. Ltd., Tokyo, Japan.
[16] V. Kayargadde and J.-B. Martens, "Perceptual characterization of images degraded by blur and noise: model," *J. Opt. Soc. Amer. A*, vol. 13, no. 6, pp. 1178–1188, 1996.
[17] S. Bouzit and L. MacDonald, "Color difference metrics and image sharpness," in *Proc. IS&T/SID Eighth Color Imaging Conf.*, 2000, pp. 261–267.
[18] G. M. Johnson and M. D. Fairchild, "Sharpness rules," in *Proc. IS&T/SID Eighth Color Imaging Conf.*, 2000, pp. 24–30.
[19] W. K. Pratt, *Digital Image Processing*. New York: Wiley, 1978.
[20] X. Zhang and B. A. Wandell, "Color image fidelity metrics evaluated using image distortion maps," *Signal Processing*, vol. 70, no. 3, pp. 201–214, 1998.

Philippe Longère received the M.S. degree in applied mathematics and the Ph.D. degree in computer science/image processing from the University Jean-Monnet, Saint-Etienne, France.

He is a currenlty a Research Scientist with m-Pixel, Sophia-Antipoles, France. His current research interests include color image processing and digital imaging.

Xuemei Zhang received the M.S. degree in statistics and the Ph.D. degree in psychology from Stanford University, Stanford, CA.

She is currently a Color Scientist with Agilent Technologies Laboratories. Her current research interests include perceptual image quality studies and algorithm development for digital imaging applications.

Peter B. Delahunt received the M.A. and Ph.D. degrees in psychology from the University of California, Santa Barbara, in 1998 and 2001, respectively.

He is currently a Post-Doctoral Fellow with the University of California Davis Medical Center, Sacramento, CA. His current research interests include human color vision.

David H. Brainard received the M.S. degree in electrical engineering and the Ph.D. degree in psychology from Stanford University, Stanford, CA, in 1989.

He is currently a Professor of Psychology with the University of Pennsylvania, Philadelphia. His current research interests include both human color vision and algorithms for processing digital color images.