EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

A no-reference metric for demosaicing artifacts that fits psycho-visual experiments

Francesca Gasparini[1*], Fabrizio Marini[1], Raimondo Schettini[1] and Mirko Guarnera[2]

Abstract

The present work concerns the analysis of how demosaicing artifacts affect image quality and proposes a novel no-reference metric for their quantification. This metric that fits the psycho-visual data obtained by an experiment analyzes the perceived distortions produced by demosaicing algorithms. The demosaicing operation consists of a combination of color interpolation (CI) and anti-aliasing (AA) algorithms and converts a raw image acquired with a single sensor array, overlaid with a color filter array, into a full-color image. The most prominent artifact generated by demosaicing algorithms is called zipper. The zipper artifact is characterized by segments (zips) with an On–Off pattern. We perform psycho-visual experiments on a dataset of images that covers nine different degrees of distortions, obtained using three CI algorithms combined with two AA algorithms. We then propose our no-reference metric based on measures of blurriness, chromatic and achromatic distortions to fit the psycho-visual data. With this metric demosaicing algorithms could be evaluated and compared.

Introduction

Image quality is difficult to assess correctly for a number of reasons [1]. Firstly, image quality is perceptual by nature. This makes it hard to measure it in a standardized way and allows for personal preferences. Secondly, it can vary widely between different application domains. Due to its perceptual nature, image quality should be evaluated through a subjective assessment, and quality metrics should be designed to fit quality judgments collected by psycho-visual experiments. The efficiency of studies that involve people's judgments is very low compared to a computerized objective study. Nevertheless, to validate automated approaches, psycho-visual scaling studies are insurmountable for image quality research [2,3]. We are interested in developing a pool of no-reference (NR) metrics to automatically assess the performance of the algorithms composing the image generation pipeline of digital cameras. In particular, we are interested in defining these metrics so that they fit the psycho-visual data. A general three-step approach to design and develop these types of metrics has been given by Bartleson [4]:

1. Identification of perceptual dimensions (attributes) of quality.
2. Determination of relationships between attribute scale values and objective, image based measures.
3. Combination of attribute scale values to predict overall image quality.

To define a no-reference image quality metric is therefore needed to design a good psycho-visual experiment. Ideally, we should be able to generate a dataset of distorted images where the distortion can be controlled by a proper defect-generating process. In this way the collected data can be easily related to the considered distortion. In particular, what we would like to obtain is a monotone behavior of the perceived quality with respect to the increase of the distortion.

Several kinds of defects can affect digital images. They can be roughly divided into [5]:

- physical defects, such as out of focus, motion blur, noise, etc.
- digital defects introduced by the processing pipeline, such as demosaicing, compression, etc.

For physical defects the procedure adopted to generate the distorted images used within the experiments could be a simulation of the physical process, while in the case

*Correspondence: gasparini@disco.unimib.it
[1]Department of Informatics, Systems and Communication, viale Sarca 336, University of Milano-Bicocca, 20126 Milano, Italy
Full list of author information is available at the end of the article

Springer

© 2012 Gasparini et al.; licensee Springer. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

of digital defects the procedure should apply the corresponding algorithm(s) within the pipeline. Note that each of these distortion processes can vary with respect to one or more parameters.

Within this context, in this paper we address the problem of how demosaicing artifacts affect image quality. The demosaicing operation converts a raw image acquired with a single sensor array, overlaid with a color filter array, into a full-color image. The most prominent artifact generated by demosaicing algorithms is called zipper. The zipper artifact is characterized by segments (zips) with an On–Off pattern.

The quality of rendered images depends on the perception of the zipper artifact that can also affect the sharpness. The perception of this artifact also depends on image content.

We here propose a no-reference metric to assess image quality in case of demosaicing artifact that combines measures of blurriness (intended as lack of sharpness), chromatic and achromatic distortions and fits the psycho-visual data. Several full-reference metrics exist for this kind of artifact [6], while the literature is poor in no-reference ones. Some no-reference sharpness metrics [7,8] could be adopted, but they can not take into account typical chromatic and achromatic zipper effects. Liu et al. [9] have recently presented a no-reference method for CFA demosaicing based on double interpolation and have evaluated several demosaicing algorithms. However this metric has not been correlated with psycho-visual experiments.

In this work we have generated a dataset with different degrees of zipper artifacts by applying a combination of three different CI algorithms with two AA algorithms. These algorithms have been applied to a set of reference images having different visual contents. More demosaicing and/or anti aliasing algorithms could have been used. However lengthy psycho-visual tests are not reliable, and we have preferred to not reduce the number of test images.

This paper is organized as follows. In Demosaicing section we briefly describe the demosaicing process, while in Psycho-visual setup section we describe how we have generated the dataset utilized during our tests and the psycho-visual experiments that we have conducted to rank the chosen algorithms. From the analysis of the experimental data (detailed in Data analysis section), we propose our novel no-reference metric, described in No-reference metric for Demosaicing section, based on measures of blurriness, chromatic and achromatic distortions. In Metric parameter estimation section we report details of the regression we have proposed to fit the subjective data and we compare our metric with a reference one [9]. All the psycho-visual data presented and the corresponding distorted images are available at http://www.ivl.disco.unimib.it/. Finally, in Section Methods we report details on the testing methodology adopted here.

Demosaicing

To produce a color image there should be at least three color samples at each pixel location. The more expensive solution consists in using a color filter in front of each sensor, generating three full-channel color images. Thus, many modern cameras use a color filter array (CFA) in front of the sensor so that only one color is measured at each pixel. This means that to reconstruct the full-resolution image, the missing two color values at each pixel should be estimated. This process, known as demosaicing [10] is generally composed of a CI algorithm followed by an AA algorithm to reduce possible artifacts. Among various CFA patterns, the Bayer pattern was the most popular choice [11]. The Bayer array measures the green image on a quincunx grid and the red and blue images on rectangular grids, obtaining 1/2 of the pixels for the green channel, and 1/4 for both the blue and the red channels, as depicted in Figure 1.

The most prominent artifact generated by demosaicing algorithms is called zipper. The zipper effect refers to abrupt or unnatural changes of color differences between neighboring pixels, manifested as an "On–Off" pattern [6]. In Figure 2 an example of an original image and two different demosaiced versions are reported. As can be seen from Figure 2b, where a typical example of demosaiced image is shown, the zipper artifacts are both chromatic and achromatic. On the other hand, demosaicing algorithms that try to mitigate this On–Off pattern, significantly blur the image (Figure 2c).

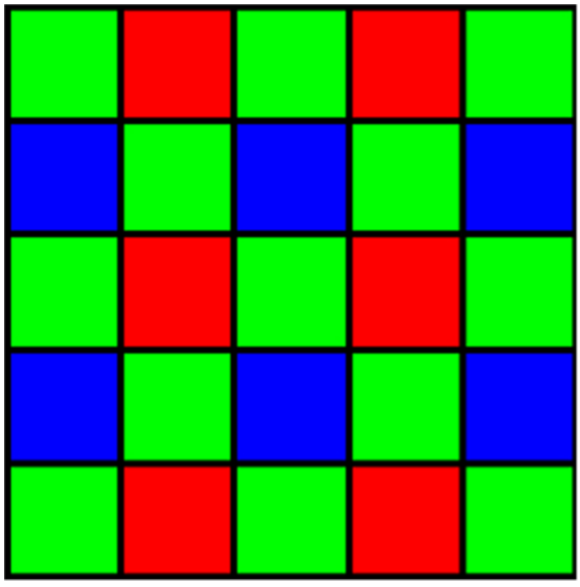

Figure 1 Bayer pattern array. The array of filters of the Bayer pattern.

Figure 2 Original image and two different demosaiced versions. (**a**) Original image. (**b**) An example of demosaicing: the artifacts introduced can be distinguished into achromatic and chromatic zipper. (**c**) A different demosaicing: the image is visibly blurred to mitigate the On–Off pattern.

Several algorithms for demosaicing were developed in the literature [12-17], and some of them are proprietary. A survey of these methods was presented by Li et al. [18]. Several methods deal with content adaptive demosaicing, based on an edge detection mechanism [19-21]. Recently Rehman and Shao [22] have presented a demosaicing method using optimised filters, based on a training process and well-defined content classification.

We have here considered nine different demosaicing algorithms obtained combining three CI algorithms with two anti aliasing (AA) algorithms.

The three CI algorithms adopted here are:

- Bilinear interpolation [18]: it is the simplest demosaicing algorithm and acts as a benchmark; the missing values on the three channels are computed by linear interpolation independently.
- ST1: proposed by Smith [23], it performs an isotropic interpolation that includes a non-linear step that minimizes the energy of aliasing artifacts.
- ST2: proposed by Guarnera et al. [24], it uses an elliptic shaped Gaussian kernel to interpolate data, according to the gradient information to better exploit spatial correlation. The authors also included an enhancement step to restore the lost high frequencies.

For what concerns the AA algorithms, we have here considered:

- an algorithm authored by Freeman [14] that suppresses demosaicing artifacts by applying a median filtering to the chrominance channels (R-G) and (B-G) to support the reconstruction of the R and B channels. The red and blue values estimated from the median filtered are used only at pixels where there is no R or B sensor value directly available.
- an algorithm authored by Lu and Tan [6] that proposes an AA step to extend Freeman's median filtering method by lifting the constraint of keeping the original CFA-sampled values intact.

The nine combinations of these algorithms (summarized in Table 1 produce different levels of the typical demosaicing distortions. The choice of these algorithms does not affect the effectiveness of the proposed methodology.

Psycho-visual setup

Testing dataset

To perform the subjective data analysis described in this paper we have generated a data set of distorted images (which we have called Zipper database) starting from the 24 images of the Kodak photoCD pcd0992 database available at http://r0k.us/graphics/kodak/. We have created the mosaiced images by deleting two of the three RGB values at each pixel of the full-color images, and then we have demosaiced them with the nine algorithms of Table 1. The database is therefore formed by a total of (24 images $\times$ 9 demosaicing methods =) 216 images. The image testing database has been created to satisfy a good compromise between the number of distortions and the number of different visual contents, keeping in mind that psychovisual sessions should be limited in time to be reliable. In our work we evaluate the visual impact of the artifacts

Table 1 Demosaicing algorithms considered

	Algorithm	Color interpolation (CI)	Anti-aliasing (AA)
1	bi	Bilinear	None
2	bifree	Bilinear	Freeman
3	bilu	Bilinear	Lu
4	ST1	ST1	None
5	ST1free	ST1	Freeman
6	ST1lu	ST1	Lu
7	ST2	ST2	None
8	ST2free	ST2	Freeman
9	ST2Lu	ST2	Lu

The nine demosaicing algorithms adopted to obtain the dataset of distorted images.

generated by demosaicing methods, and do not perform a quality evaluation of the algorithms themselves.

Testing methodologies

For the quality analysis of the images we adopted two different test methods: single stimulus method (1S), and double stimulus method (2S) [3].

Our goal was to evaluate the perceived quality of the rendered images; for this reason we have chosen to set up a single-stimulus test as our primary source of psycho-visual data, but we were also interested in gathering as much data as possible from the viewers, so we have also conducted a double-stimulus test. We followed Sheikh et al. [25] in setting up our tests by including the original images in both tests and calculating the Difference Score (DS) as the difference between the scores of the original and the distorted image. This way we have obtained different data from different setups with the same unit of measure. In the case of the 1S method, all the images (rendered images and the original one) are individually shown. While in the 2S method, the reference image (original image) is shown together with each of its rendered versions. The 1S method can thus be considered as an approximation of the 2S one, as the original image is evaluated only once. The fundamental difference between these two methods is that the 2S one uses an explicit reference, while the 1S one does not use any explicit reference.

To perform the psycho-visual tests, the images that have to be judged to obtain a quality rank were shown on a web-based interface (Figure 3). A Javascript slider assigning a quality score was used. The workstations adopted were placed in an office environment with normal indoor illumination levels [25,26].

We used five 19-in. CRT COMPAQ S9500 display monitors:

- All the monitors were calibrated with a colorimeter (D65, gamma 2.2).
- Their resolution is 1600 × 1200 pixels, which corresponds to 110 dpi (using 18 in. as the physical diagonal of the screen as indicated by the manufacturer of the monitors)
- The ambient light levels (a typical office illumination) were maintained constant between the different sessions. There were no reflections on the screens.
- The distance between the observer and the monitors was about 60 cm (corresponding to about 46 pixels per degree of visual angle).
- The refresh rate of the monitors was 75 Hz.

In all our experiments distorted images are shown in random order, different for each subject. In the case of the 2S method the relative position of the original with respect to its distorted version is random in the pair shown.

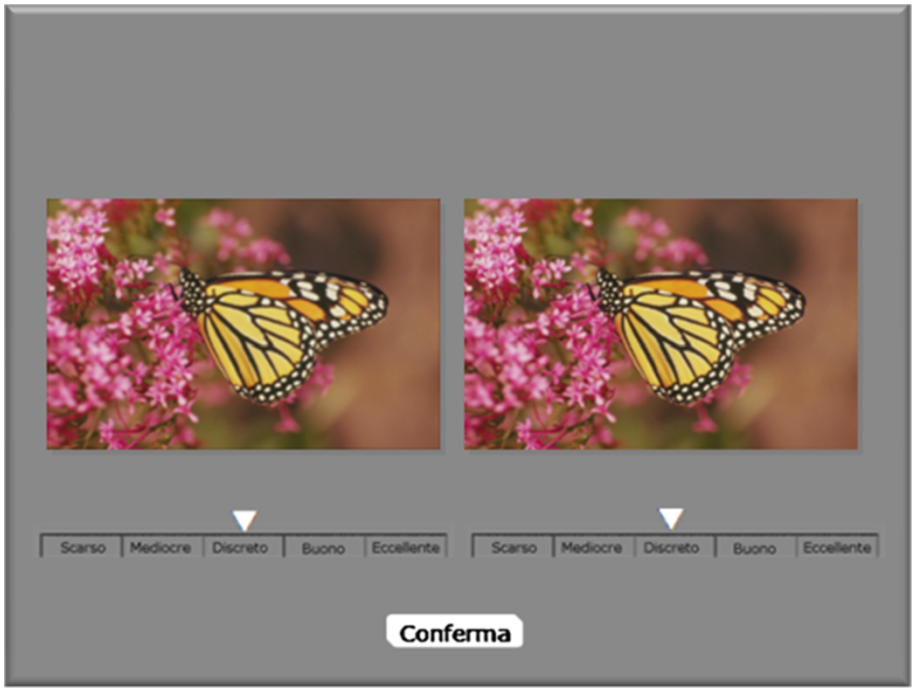

Figure 3 Web interface DS. The web interface used during the Double Stimulus tests.

The panel of subjects involved in this study was recruited from the Psychology Department. The subject pool consisted of students inexperienced with image quality assessment and image impairments. The total number of subjects involved in our experiments is 39, divided into three groups as follows: 9 subjects involved in tuning experiments, and 30 subjects involved in 1S and 2S experiments, 15 for each test group.

Psycho-visual experiments

In our experiments for the collection of subjective data, we performed three main sessions: a tuning session (where we verified the test efficacy and the best way to perform the experiment), a preliminary session (where we trained the observers about the nature and the range of the distortion) and a final test session. Details of the tuning and of the preliminary sessions are reported in Section Methods. For the test session we used 10 images from the 24 of the original database, together with their corresponding 9 distorted versions, (for a total of 100 images). The 10 images chosen for this session are shown in Figure 4. Note that we had to keep the number of analyzed images limited to 100, since subjects can pay attention only for up to 30 min. After this time their judgments are no longer reliable [3]). The number of test images is however aligned with what is done in the literature. In the work of Nyman et al. [27] for example, 9 image processing pipes applied to 8 image contents (for a total of $9 \times 8 = 72$ test images) were evaluated with a psycho-visual experiment involving 14 test subjects. In other works that involve psycho-visual experiments, the number of original images considered is even lower, four images each printed on 15 different papers [28], or five images each with 15 different levels of sharpness [29]. The greater the number of algorithms/processing to be evaluated, the lower the number of original images that can be considered to keep the time of the experiment reasonable.

Data processing

As the different algorithms considered produce different levels of the typical demosaicing defects (chromatic and achromatic zipper, blur) we analyzed the subjective

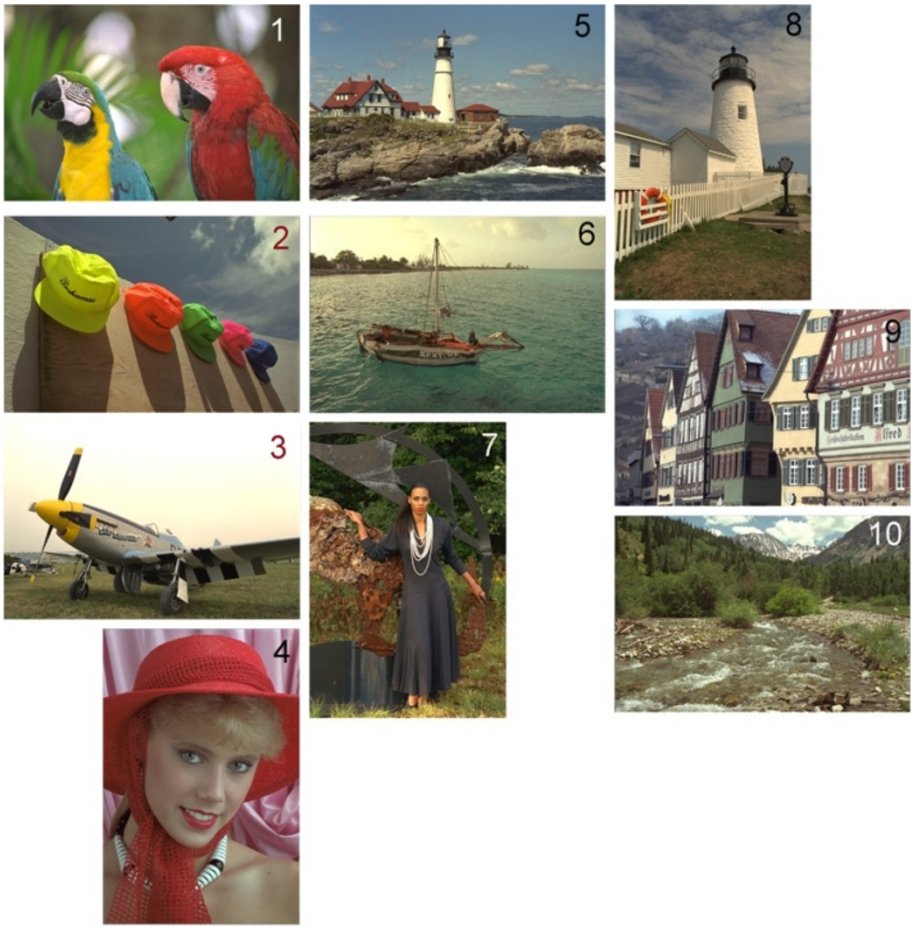

Figure 4 Used dataset. The ten original images utilized during the test session.

evaluation of these defects through the subjective rank of the algorithms.

The data processing here described is applied for both the test methods (1S, 2S) adopted for collecting the experimental data. For each subject j-th and distorted image i-th we evaluated the perceptual distance between original and distorted images in terms of difference of assigned scores (Difference Score, DS):

$$\mathrm{DS}_{ij} = \mathrm{So}_{ij} - \mathrm{Sd}_{ij} \tag{1}$$

where Sd_{ij} represents the score assigned by the j-th subject to the i-th distorted image, while So_{ij} the score of the reference image corresponding to the i-th distorted image; $j = 1, \ldots, J$, denotes subjects belonging to the group of J individuals, and $i = 1, \ldots, S \times T$ denotes the distorted image, with S number of reference images, and T number of algorithms to be evaluated. For each subject we evaluated the standard-DS_{ij} ($z\mathrm{DS}_{ij}$), a DS distribution normalized with respect to the subject [25], as:

$$z\mathrm{DS}_{ij} = \frac{\left(\mathrm{DS}_{ij} - M_j\right)}{\sqrt{V_j}} \tag{2}$$

where $M_j = \frac{1}{S\times T}\sum_i^{S\times T} \mathrm{DS}_{ij}$, $V_j = \frac{1}{S\times T}\sum_i^{S\times T}\left(\mathrm{DS}_{ij} - M_j\right)^2$ are respectively the mean and variance of DS_{ij} with respect to the j-th subject. For each algorithm $t \in T$, we evaluated the final score R_t by summing $z\mathrm{DS}_{ij}$ of Equation 2 over the subjects $j \in J$, and on the reference images $i \in S$.

$$R_t = \frac{1}{J \times S}\sum_{j\in J}\sum_{i\in S} z\mathrm{DS}_{ij} \tag{3}$$

The rank of the algorithms is then obtained sorting these final scores. We also calculated the rank of the algorithms starting from the median with respect to subject $j \in J$ and reference images $i \in S$.

$$\mathrm{MR}_t = median(z\mathrm{DS}_{ij}) \tag{4}$$

Data analysis

In analyzing distorted images supposed to be worse than the original, we expect all the DS values (distance between the scores of the original image and the rendered image) to be positive. It happened sometimes in our experiments that distorted images were judged better than the corresponding original. This phenomenon is called inversion. We have decided to maintain all the inversions. The reasons for this decision are detailed in Section Methods. We want to emphasize that with this data analysis we are not evaluating the performance of the algorithms, but instead we are interested in highlighting the major effects that influence the subjective evaluations of the perceived quality of demosaiced images. The final goal is to identify the significant features to be used in a proper metric so that it can be able to reproduce the experimental data. In Figure 5, the rank of the nine demosaicing algorithms obtained combining the three CI algorithms with the two AA algorithms listed in Table 1 are reported for both the 1S and the 2S experiments. Figure 5a and Figure 5b show the ranks of the 2S experiment using respectively the mean R_t and the median MR_t as a central tendency indicator. The coherence between these two ranks confirms the stability of the results. In Figure 5c and Figure 5d, the same data are reported for the 1S experiment. In Figure 6a comparison of the two experiments is reported. The solid line refers to the 1S experiment, while the dotted line refers to the 2S experiment.

As a preliminary step, we have grouped the 9 demosaicing methods into triplets, with respect to the CI algorithm applied.

As a general consideration, CI algorithms alone (i.e. bilinear, ST1 and ST2) were judged worse than their corresponding versions coupled with any of the AA algorithms considered. With respect to the CI approach, the ST2 method (coupled with any AA algorithm) is always preferred as it produces sharper images. This behavior is due to the explicit boosting introduced by the authors to restore the lost high frequencies. These results confirm that sharpness plays an important role in influencing image quality judgments [30,31]. 1S tests are less precise than 2S tests because the reference image is shown only once, and the comparison between distorted images and reference ones is more difficult. Were this the only difference between the two tests, we would not expect significant changes in the algorithm ranks. This assumption was not fully verified in our experiments. This discrepancy is also due to the effect of the perceived sharpness on image quality, which is more evident in 2S tests due to the direct comparison with the reference images. The AA algorithms considered have influenced the image sharpness at different degrees. In particular the Freeman algorithm produces a sharper image, while the Lu algorithm makes the images more blurred. This phenomenon is more evident when these AA algorithms are coupled with the basic CI method (bilinear interpolation) as shown in Figure 7. As a consequence, the rank positions of the algorithms labeled as bifree (algorithm 2) and st2free (algorithm 8) are swapped from the 2S to the 1S experiment with respect to the corresponding bilu (algorithm 3) and st2lu (algorithm 9) as shown in Figure 7.

Image content

We have analyzed the experimental data to investigate the cross-talks between the zipper artifacts introduced by the CI process and the image content. In Figure 4 the

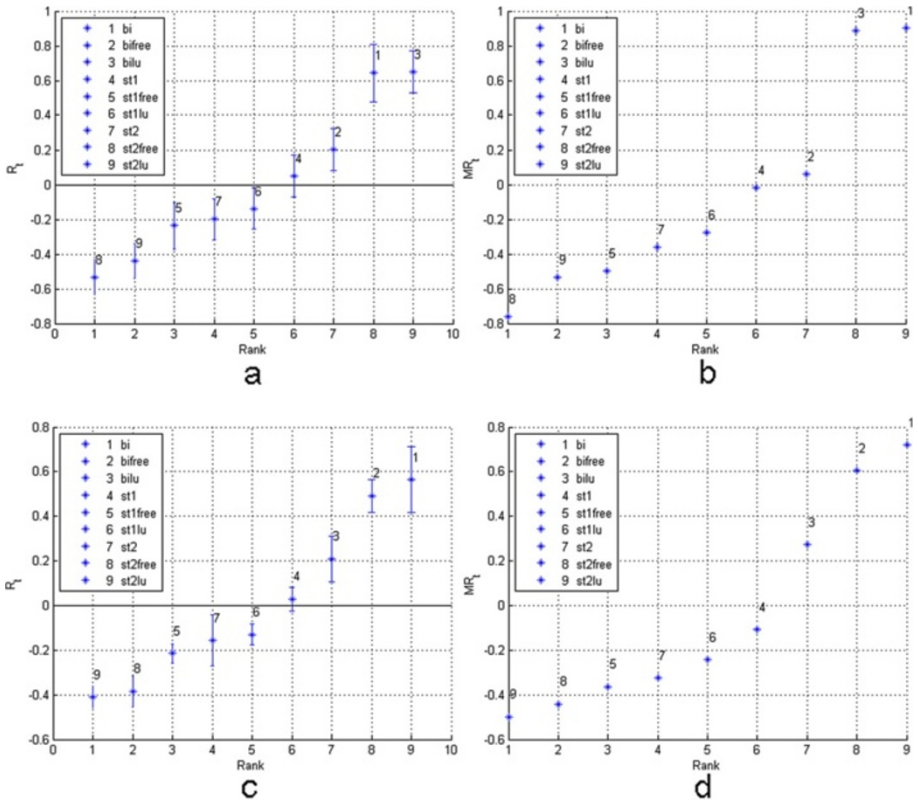

Figure 5 Psycho-visual results. (**a**) Algorithm ranks in terms of final score R_t in the 2S experiment. **(b)** The 2S rank resulting from using the median MR_t as a central tendency indicator [25]. **(c)** Algorithm ranks in terms of R_t in the 1S experiment. **(d)** The 1S rank resulting using the median MR_t as a central tendency indicator.

10 images used in our tests are listed from 1 to 10 with increasing value of visual complexity as obtained applying the complexity index described in [32].

To better understand how the visual content influences the psycho-visual data, we have collected the subjective score (Score_i) for each of the ($S = 10$) test images and for each of the ($T = 9$) demosaicing algorithms. Summing the $z\text{DS}_{ij}$ of Equation 3 over the subjects $j \in J$ we obtain:

$$\text{Score}_i = \frac{1}{J} \sum_{j \in J} z\text{DS}_{ij} \tag{5}$$

with $i = 1, \ldots, S \times T$. The Score_i are reported for both 2S and 1S experiment in Figure 8, where the layout of Figure 4 is maintained.

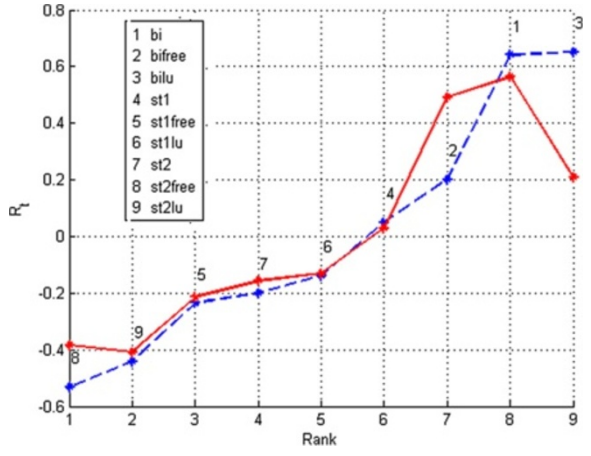

Figure 6 Comparison 1S 2S. Comparison between R_t values of the 1S experiment (red solid line) and the R_t values of the 2S experiment (blue dotted line).

Each subplot reports the experimental Score_i corresponding to the nine distortions applied to each image. These scores are grouped into triplets with respect to the CI method (bilinear + three AA, ST1 + three AA, and ST2 + three AA).

We can notice that images with a comparable level of details share common patterns in their scores. In particular, when the achromatic zipper (mainly produced by the Freeman AA algorithm) is combined with middle-high frequency content (roughly second and third column of Figure 4), not only the contrast of the zipper highlights the edges, but also the middle-high frequency content masks the On–Off pattern. This combined effect results in a sharper appearance of the image; this is more evident

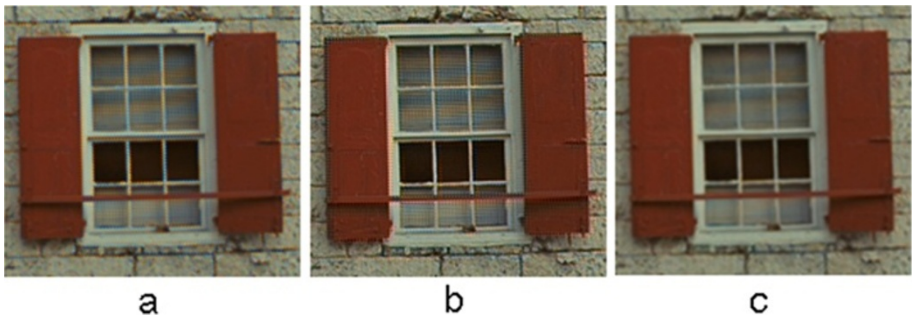

Figure 7 Example of demosaicing. Detail of an image rendered with different algorithms (**a**) Bilinear (CI) (**b**) Bilinear (CI) + Freeman (AA) (**c**) Bilinear (CI) + Lu (AA).

when the images are directly compared with the reference, as in the 2S test. This behavior is related to the texture masking effect of the human visual system [33]. From the point of view of the algorithm ranks (Figure 8) these considerations are confirmed by the good performance on these images obtained using the Freeman AA with respect

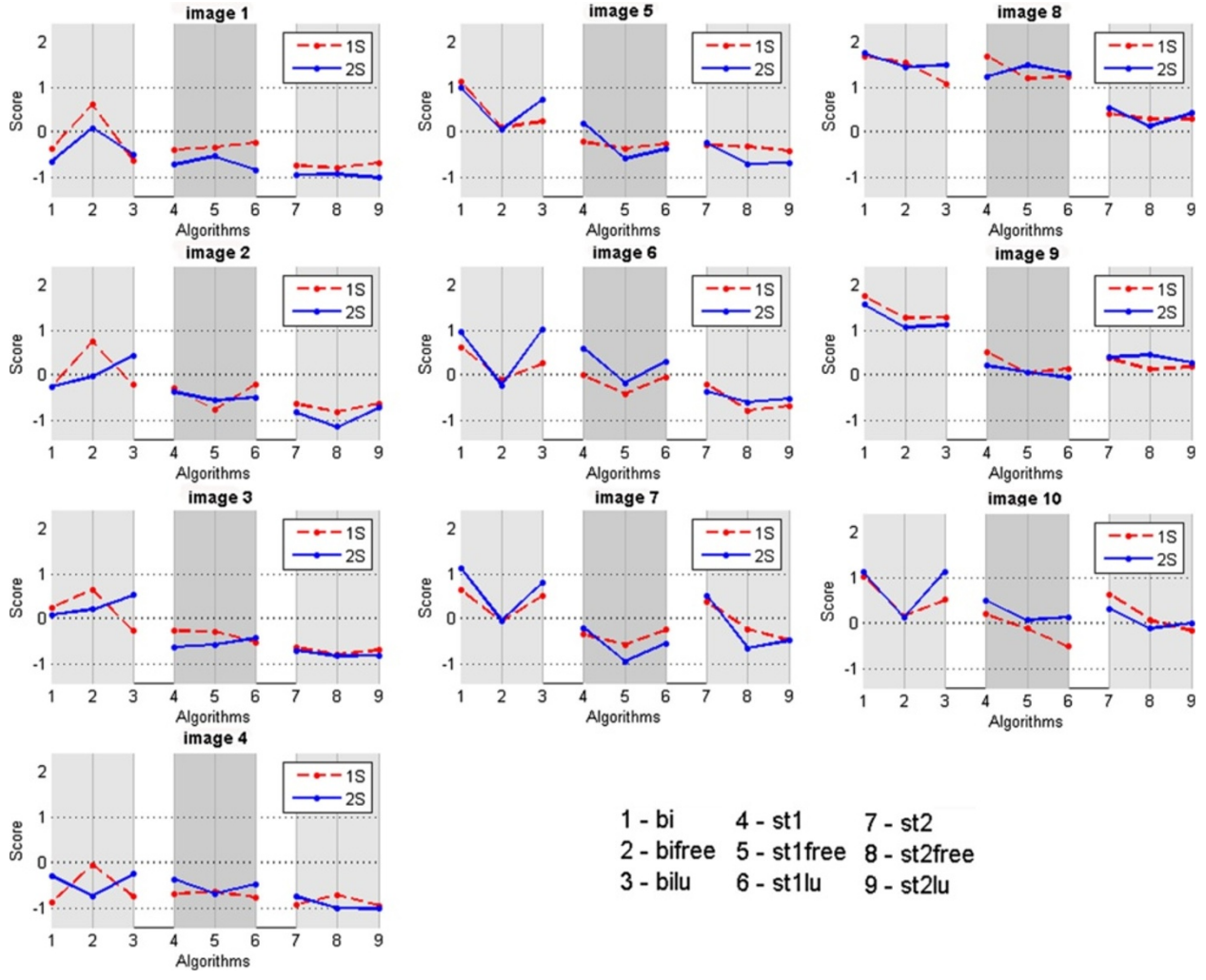

Figure 8 Subjective test data for each image. Subjective test data. Each subplot refers to the corresponding image of Figure 4. The scores are grouped in triplets, corresponding to each of the three CI methods coupled with the three different AA strategies. For instance, the first triplet corresponds to algorithms 1, 2 and 3, i.e. bilinear interpolation with no AA, Freeman AA and Lu AA respectively.

to each triplets of CI algorithms (algorithms number 2, 5, 8). On the other hand, when the algorithms that produce this achromatic distortion are applied to images with a low frequency content (first column), the high contrast of the zipper pattern and its On–Off structure remain visible. In fact, the evaluation of the CI algorithms coupled with the Freeman AA in the case of low frequency content is worse than in the case of higher frequency contents, especially in the case of the 1S experiment where the sharpness is less perceived.

Images 9 and 10 are characterized by a texture with a near-Nyquist frequency as shown in Figure 9, where the distortions due to the aliasing are evident. These images have suffered from very strong distortion after the CI process and thus their subjective scores could have been reduced by the near-Nyquist artifacts.

For what concerns the chromatic zipper, the behavior is simpler. This artifact is more visible as the number of edge pixels in the image increases, and it seems to be immune to masking effects. For this reason we chose to discriminate between chromatic and achromatic distortion.

No-reference metric for demosaicing

The data analysis confirms that the perceptual quality of demosaiced images depends on sharpness, and on chromatic and achromatic zipper. For this reason we have decided to define our no-reference metric considering the following three aspects separately:

Blur as index of lack of sharpness. The corresponding measure is indicated as B in what follows.
Chromatic zipper distortion (measure indicated as CD)
Achromatic zipper distortion (measure indicated as AcD)

Thus, the demosaicing metric DM that we have developed is composed of three properly scaled terms, corresponding to these three aspects:

$$\mathrm{DM} = B + \mathrm{CD} + \mathrm{AcD} \tag{6}$$

Figure 9 Near Nyquist artifacts. Details of the images 9 and 10 of Figure 8 with near-Nyquist frequency content.

We chose a sum expression because when one of these terms is significantly high, the others are less significant. This consideration arises from the experimental evidence of the behavior of different demosaicing algorithms. A strong low pass filtering adopted to reduce the zips produces a blurred image, and thus in this case the blur measure B is dominant with respect to the others. In case of more conservative filtering, the image sharpness is preserved, but the zips still remain as a defect. Different CI algorithms produce zips with different levels of saturation, ranging from achromatic to highly saturated zips.

Blur

The blur in an image is due to the attenuation of the high spatial frequencies. Blur is the typical artifact in out-of-focus shots, but it may also be caused by the relative movement between camera and subject (motion blur), and by the encoder (compression blur). In the context of CI artifacts, blurriness is due to an excessive low pass filtering of the AA algorithm. Marziliano et al. [7] present a blind (no-reference) blur metric that is based on measuring the average spread of the vertical edges. They define the edge spread as the distance between the local minima (p_1) and the local maxima (p_2) nearest to the edge along the gradient direction (Figure 10). The edge spread was used to predict the quality of jpeg2000 compressed images and has shown to be consistent with the observers' ratings obtained in subjective experiments. We use as blur indicator, the average edge spread of the image **Es**, evaluated as follows:

$$\mathbf{Es} = \frac{1}{N\mathrm{Edge}} \sum_{e \in \mathrm{Edge}} \mathrm{dist}_4(p_1, p_2) \tag{7}$$

where Edge is the set of edge pixels of the image and NEdges is the number of these edge pixels. We chose to estimate the edge spread by searching around the edge in four directions (indicated with dist_4 in 7): horizontal, vertical, +45° and −45°.

Chromatic and achromatic zipper

The zipper pattern detection was carried out as follows. On the gray-scale image, we computed the gradient magnitude in both directions with the following convolution kernels:

$$\begin{aligned} V &= \begin{bmatrix} -1 & 1 \end{bmatrix} \\ H &= \begin{bmatrix} -1 & 1 \end{bmatrix}^T \end{aligned} \tag{8}$$

The two gradient maps, G_x and G_y (horizontal and vertical), are treated separately to detect zipper segments. Working on the horizontal direction, we first compute the gradient sign map by quantizing the gradient magnitude

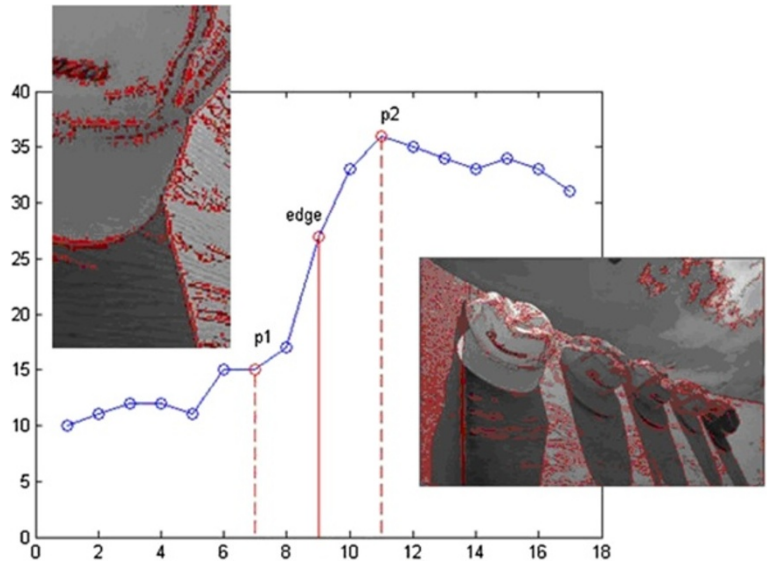

Figure 10 Edge spread measure. Edge spread defined as the distance between the local minima (p_1) and the local maxima (p_2) nearest to the edge.

as follows (the same process is extended to the vertical direction):

$$SignMap_x(x,y) = \begin{cases} 2 & \text{if } G_x(x,y) < 0 \\ 1 & \text{if } G_x(x,y) > 0 \\ 0 & \text{if } G_x(x,y) = 0 \end{cases} \tag{9}$$

Thus, a zipper segment (which is an On–Off pattern) is characterized in the sign map by a sequence of alternating 2s and 1s (see Figure 11). The number and the extension of zips is not sufficient to quantify the perceived quality of a CI algorithm. In fact, some zipper pixels are more visible than others (see Figure 11c).

To evaluate the visibility of the pixels belonging to the zipper segments, we compute $\mathrm{DL}(x,y)$ and $\mathrm{DC}(x,y)$ distances between adjacent pixels in zipper segments, starting from the CIE-94 definitions [34]:

$$\mathrm{DL}(x,y) = \left(\left(\Delta L^*(x,y)\right)^2\right)^{\frac{1}{2}} \tag{10}$$

$$\mathrm{DC}(x,y) = \left(\left(\Delta C^*(x,y)/S_c\right)^2 + \left(\Delta H^*(x,y)/S_H\right)^2\right)^{\frac{1}{2}}$$

where[1]

$$\Delta L^*(x,y) = L^*(x,y) - L^*(x,y-1) \tag{11}$$

$$\Delta C^*(x,y) = ((a^*(x,y))^2 + (b^*(x,y))^2)^{\frac{1}{2}} - ((a^*(x,y-1))^2 + (b^*(x,y-1))^2)^{\frac{1}{2}}$$

$$\Delta H^*(x,y) = ((\Delta E_{76}(x,y))^2 - (\Delta L^*(x,y))^2 - (\Delta C^*(x,y))^2).$$

We calculated the median of $\mathrm{DL}(x,y)$ with respect to the whole set of zipper segments in both directions and averaged them. We performed the same calculations for $DC(x,y)$, obtaining two indicators labeled as **DL** and **DC**

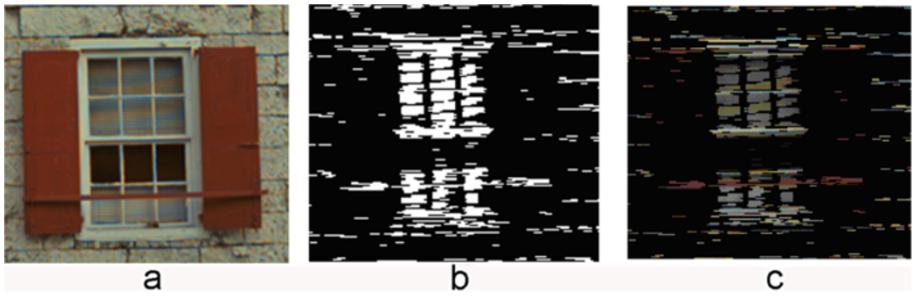

Figure 11 Zipper maps. (**a**) Detail of an image rendered with bilinear interpolation. (**b**) Horizontal zipper map. (**c**) The original image masked with the horizontal zipper map.

in what follows. These two indicators, together with the average edge spread (**Es**) and the percentage of zipper pixel in the image (ZpA), were used to calculate the overall metric.

Metric parameter estimation

Starting from the the blur and zipper pattern analysis described in the two previous subsections, our demosaicing metric (6) can be rewritten as:

$$\mathrm{DM} = w_B \times \mathbf{Es} \times + w_C \times \mathbf{DC} + w_L \times e^{\mathbf{DL}-\mathbf{DC}} \times ZpA \quad (12)$$

Algorithms that reduce aliasing tend also to desaturate the zips, increasing the coherence between channels. This effect produces achromatic zips, where **DL** exceeds **DC**. w_B, w_C and w_L are weights chosen using an exhaustive search algorithm, so that our metric can fit as better as possible the algorithms' rank produced by the psycho-visual experiments. To this end, we have applied the proposed metric to the images in the Zipper Database, and then we have calculated the average metric scores of the nine algorithms. We have performed a regression analysis to find the best fit between the average values given by our measure and the average subjective responses R_t for both the 2S and 1S data. In Figure 12 we have reported the logistic regressions that we have obtained respectively for 1S (Figure 12a) and 2S (Figure 12b) experiments. The weight sets we have found for both the experiments are unique. In theory more weight sets could have the same quality performance. In this case to choose the set to be used we would follow the experimental procedure described in [27,28], to order the subjective importance of visual attributes. We have also compared our metric with the DIPSNR no-reference metric proposed by Liu et al. [9]. Figure 13 reports the linear regressions that best fit the DIPSNR values with the psycho-visual data for both the experiments (1S, Figure 13a, 2S Figure 13b), while in Figure 14 the corresponding logistic regressions are also reported. In Table 2 our metric is compared with the DIPSNR in terms of the Pearson (prediction accuracy), the Spearman (prediction monotonicity), the Kendall (rank correlation) and the MAPE (mean absolute percentage error of the prediction) coefficients.

The Pearson and Spearman coefficients above 0.98 indicate that our metric is highly accurate and monotonic. The corresponding coefficients of the DIPSNR metric are lower, as PSNR techniques are numerical measures that usually do not correlate well with perceived distortions [26].

The contribution of each term adopted in our metric, can be investigated by looking at the different values assumed by the corresponding weights w_B, w_C and w_L in Equation 12. These values are reported in Table 3. The main difference between 2S and 1S experiments is in the contribution of the sharpness. In fact, in the case of 2S experiment, where a reference image is shown, the difference in sharpness is more evident, thus the corresponding weight w_B is higher than in the 1S experiment.

Conclusions

In this work we have set up psycho-visual experiments to analyze the subjective evaluation of the artifacts introduced by the demosaicing process. To this end we

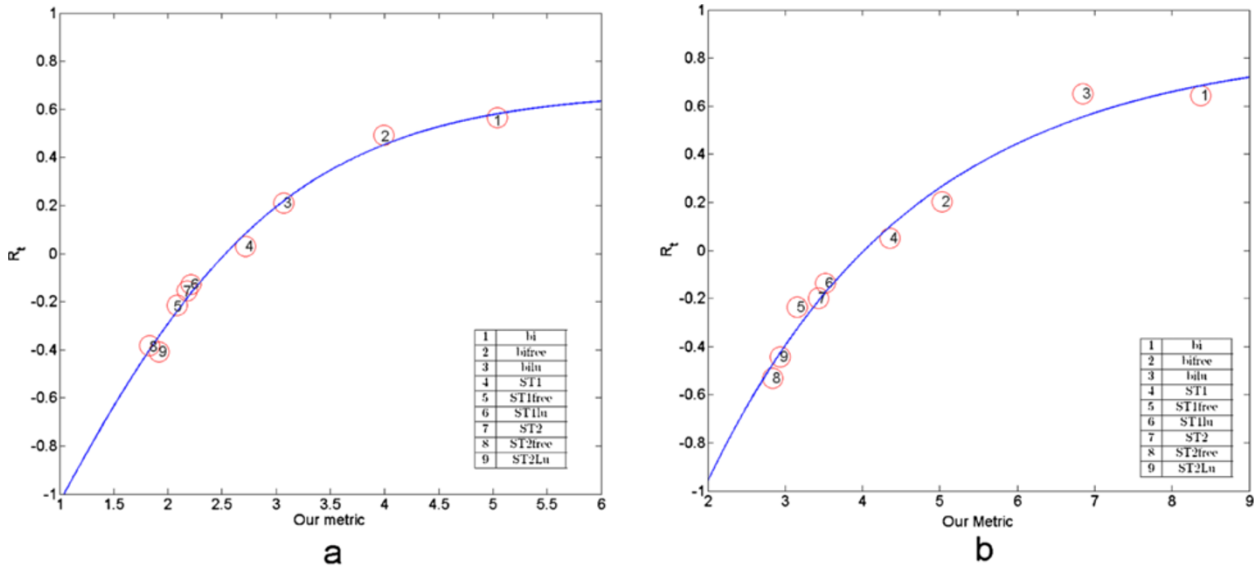

Figure 12 Our metric regressions. Logistic regression of our metric values to fit the psycho-visual data. (**a**) Single Stimulus (1S), (**b**) Double Stimulus (2S) .

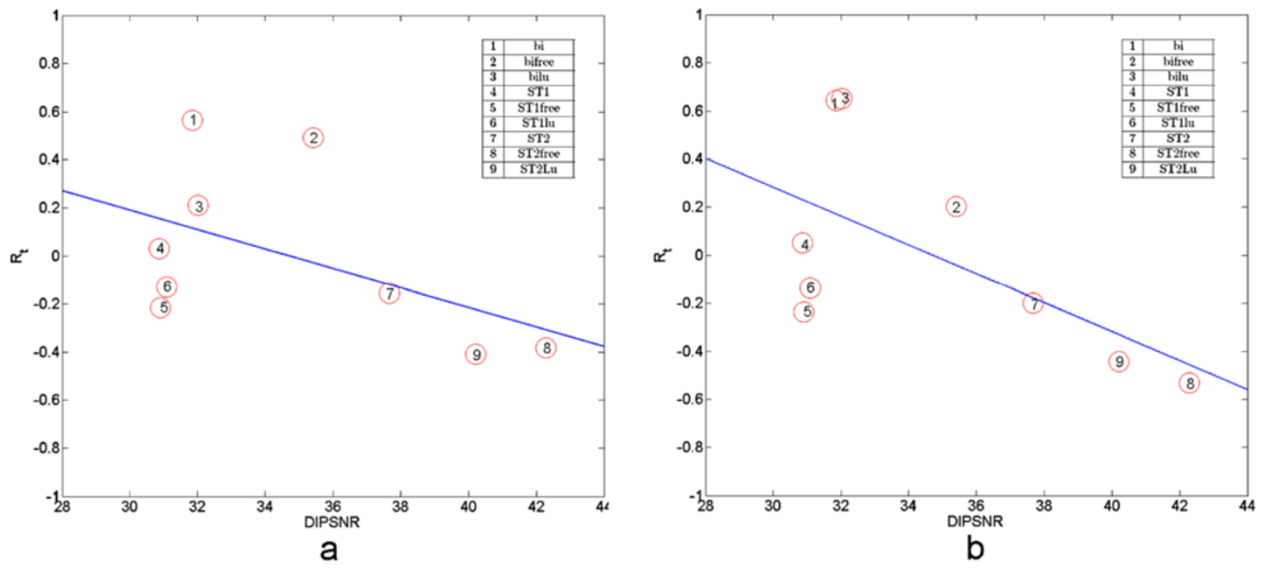

Figure 13 DIPSNR linear regressions. Linear regression of DIPSNR values that best fits the psycho-visual data. (**a**) Single Stimulus (1S), (**b**) Double Stimulus (2S).

have generated a dataset of distorted images, applying three CI algorithms combined with two AA algorithms for a total of nine different methods. From the data analysis, it emerges that the perceptual quality of demosaiced images mainly depends on perceived sharpness, and on chromatic and achromatic zipper. The perception of the defects is more evident when the rendered images are compared with the reference one (the 2S experiment), while they may be unnoticed when images are evaluated alone (the 1S experiment). We have thus defined a no-reference metric for demosaicing artifacts based on measures of blurriness, chromatic and achromatic distortions that is able to fit these experimental data for both 1S and 2S experiments. Our metric can be applied to evaluate other demosaicing methods. As a future work we plan to perform further test sessions to acquire more

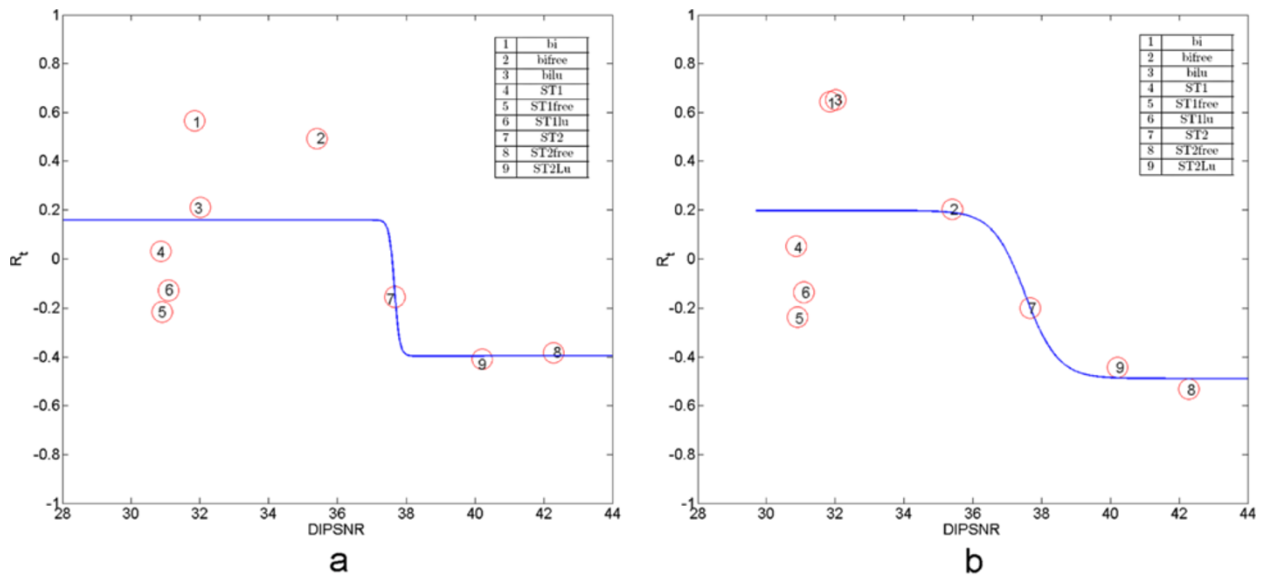

Figure 14 DIPSNR logistic regressions. Logistic regression of DIPSNR values that best fits the psycho-visual data. (**a**) Single Stimulus (1S), **(b)** Double Stimulus (2S).

Table 2 Statistic parameters

	Our metric		DIPSNR			
Experiment	1S	2S	1S		2S	
Regression	*logistic*	*logistic*	*linear*	*logistic*	*linear*	*logistic*
Pearson	0.994	0.989	0.502	0.615	0.697	0.710
Spearman	0.983	0.983	0.417	0.433	0.602	0.433
Kendall	0.944	0.944	0.287	0.333	0.458	0.333
MAPE	0.321	0.269	1.270	1.265	1.125	0.986

Comparison of the performance of our metric and the DIPSNR in case of both 1S and 2S experiments.

data to better analyze the cross-talk between distortion perception and image frequency content.

Methods

Details of experimental sessions

In our experiments for the collection of subjective data, we have performed different sessions of tests, with different goals. The three main categories of sessions are:

- Tuning session
- Preliminary sessions
- Test sessions

The total number of subjects involved in our experiments is 39, divided into three groups: (i) 9 subjects involved in tuning experiments, (ii) 15 subjects involved in the 1S experiments (both preliminary and test sessions), (iii) 15 subjects involved in the 2S experiments (both preliminary and test sessions).

Note that each subject only belongs to one group. Each subject has been individually briefed about the modality of the experiment in which he has been involved.

All the images utilized for the psycho-visual tests were cropped to fit the dimension of the screen. In particular, to avoid the undersampling of the images used in the 2S tests, we have cropped all the images to fit a 600×600 box, producing respectively images of 600×512 or 512×600. The remaining part of the box has the same color of the background (Figure 3). Each image has been cropped manually to keep the relevant part of the scene centered, to avoid interferences in the user's judgment, due to a non significant cropping.

Tuning session

Before starting the preliminary and test sessions, an initial analysis of the test structure and organization was performed to better tune the successive experiments. The 9 subjects participating in this session were not involved in other experiments. During this tuning session we verified the test efficacy and the best way to perform the experiments. In particular, we defined the best visualization time for each image or pair of images on the screen, and the maximum duration of the whole experiment for each participant. We have also collected the following considerations:

- The subjects assume and maintain the correct position and distance from the monitor for the duration of the experiment.
- 30 min is the maximum duration of the test for each subject. For longer periods attention decreases and subjects tend to get tired.
- In the case of 2S test, where the two images are compared, the sliders and the quality scales must appear contemporarily on the screen.

Regarding comments and considerations of the subjects involved in this tuning session, we have determined the minimum time of image visualization that permits an appropriate quality evaluation.

Preliminary session

During a preliminary test, each subject was implicitly trained about the nature of the distortion he was going to evaluate. In particular, he was trained about the range of the distortion intensity. These preliminary sessions were necessary to avoid this training phase during the effective test, thus conditioning the experimental results. We had preliminary sessions for all the subjects involved (except for 9 subjects involved in the tuning phase) and for each of the experiments (1S and 2S). Thus we had preliminary sessions for all the subjects involved and for each of the experiments: 2S and 1S. Four images were chosen from the entire database. The demosaicing algorithms applied to these images where the Bilinear and the ST proprietary. We have decided to apply these two algorithms because they were supposed to be the worst and the best ones. In this way the subjects experience the entire distortion range before starting the effective test.

Inversions

In analyzing distorted images supposed to be worse than the original, we expect all the DS values (distance between the scores of the original image and the rendered image) to be positive. It happened sometimes in our experiments that distorted images were judged better than the corresponding original. This phenomenon is called inversion. We define Just Noticeable Difference Threshold (JND) the threshold under which differences between distorted

Table 3 Metric weights

Weights	2S	1S
w_B	5.0	2.0
w_L	5.0	5.0
w_C	1.5	1.2

Weights for the 1S and 2S test data.

images and their original are not noticeable. Assuming that this threshold exists, the inversions can be classified into three categories:

JND inversions The subject is not able to distinguish between the original and the distorted image. The inversion is unintentional.
Preferential inversions The subject prefers the elaborated image.
Error inversions The subject does not properly use the interface and in particular, assigns a wrong value in the quality scale.

As reported in [3], the inversions are usually handled, following a standard procedure:

1. The JND threshold is estimated with a Pairwise Comparison (PC) test [35];
2. Inversions that produce values under the JND threshold (JND inversions) are taken into account in the final analysis;
3. Inversions that produce values over the JND threshold are considered as error inversions. Their absolute values are taken into account in the final analysis.

Preferential inversions

In [36], the authors report interesting considerations about preferential inversions in case the of images processed by demosaicing algorithms. They analyze the results of a Pairwise Comparison test of images processed by different demosaicing algorithms. This psycho-visual experiment demonstrates that certain algorithms produce distorted images judged better that the original. This preference is due to the apparent sharpness introduced by these algorithms. It is well known that sharpness plays an important role in the evaluation of apparent quality of digital images, [30,31]. Using a 2S method as the PC test, the original image appears blurred in comparison with the elaborated one. Not all the demosaicing algorithms analyzed in our experiment show the same sharpening behavior. As a consequence, the collected data are non-homogeneous with respect to algorithms that present different levels of preferential inversions. Applying the standard procedure, preferential inversions are not explicitly considered. These Inversions fall both in the error inversions and in the JND inversions. For this reason we have decided to maintain all the inversions. This decision requires the solution to two different problems:

- How to treat the error inversions?
  The error inversions cannot be common to different subjects. They are anomalous values with respect to the score distribution of each algorithm. We are not interested in finding the error inversions; we just would like to verify that they do not alter the data analysis. To this end we have validated the final rank of the algorithms (R_t in Equation 3 (which is a mean measure), also with the analysis of the median of the Difference Score, which is a more robust measure with respect to noise.
- How to treat the preferential inversions?
  Maintaining the preferential inversions, the DS measure cannot be further considered as a distance between the reference image and the distorted one with respect to the analyzed artifact (zipper artifact), as we have previously discussed. The influence of these inversions appears to be different in the case of 1S and 2S tests. In fact, the effect of the introduced sharpness is lower in the case of the 1S test because there is not a simultaneous comparison with the original image. Thus, the analysis of the 1S test results with respect to the 2S ones can be useful for evaluating this phenomenon.

Endnote

[1]The equations are reported only for the horizontal case. In the calculation of the differences we excluded the non-zipper pixels. ΔE_{76} is the standard Euclidean distance between the $L^*a^*b^*$ coordinate of the adjacent pixels.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
Our investigation was performed as a part of a ST Microelectronics's research contract. The authors thank ST Microelectronics for the permission to present this paper.

Author details
[1]Department of Informatics, Systems and Communication, viale Sarca 336, University of Milano-Bicocca, 20126 Milano, Italy. [2]ST Microelectronics, AST Catania Lab, Imaging Group, St.le Primosole 50, 95121 Catania, Italy.

Received: 28 October 2011 Accepted: 30 May 2012
Published: 21 June 2012

References

1. PG Engeldrum, Psychometric scaling: avoiding the pitfalls and hazards. in *IS&T's 2001 PICS Conference Proceedings*. (Montreal Quebec Canada, vol. 4, 101–107, 2001)
2. Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality. ITU-T Study Group 9 Contribution 80 (2000)
3. Recommendation 500-11, Methodology for the Subjective Assessment of the Quality for Television Pictures. ITU-R Rec. BT, 500 (2002)
4. J Bartelson, The combined influence of sharpness and graininess on the quality of colour prints. J. Photogr. Sci. **30**, 33–38 (1982)
5. R Lukac, *Single-Sensor Imaging: Methods and Applications for Digital Cameras*. (Boca Raton: CRC Press, 2008)
6. W Lu, Y Tan, Color filter array demosaicing: new method and performance measures. Image Process. IEEE. **12**, 1194–1210 (2003)
7. P Mariziliano, F Dufaux, S Winkler, T Ebrahimi, Perceptual blur and ringing metrics: application to jpeg2000. Signal Process. Image Commun. **19**, 163–172 (2004)

8. R Ferzli, L Karam, A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB). IEEE Trans. Image Process. **18**(4), 717–728 (2009)
9. Y Liu, Y Lin, S Chien, A no-reference quality evaluation method for CFA Demosaicking. Digest of Technical Papers International Conference on Consumer Electronics (ICCE). **1**, 365–3668 (2010)
10. B Gunturk, J Glotzbach, Y Altunbasak, R Schafer, R Mersereau, Demosaicking: color filter array interpolation in single chip digital cameras. IEEE Signal Process. Mag. **22**, 44–54 (2005)
11. B Bayer, Color imaging array. U.S. patent 3971065, 1976
12. DR Cok, Signal processing method and apparatus for producing interpolated chrominance values in a sampled color image signal. U.S. patent 4 642 678, 1986
13. R Kimmel, Demosaicing: image reconstruction from color CCD samples. IEEE Trans. Image Process. **8**, 548–258 (1999)
14. TW Freeman, Median filter for reconstructing missing color samples. U.S. Patent 4724395, 1998
15. DH Brainard, D Sherman, Reconstructing image from trichromatic samples: from basic research to practical applications. in *IS&T/SID Color Imaging Conference*. (Scottsdale, AZ, 4–10 1995)
16. B Leung, G Jeon, E Dubois, Least-squares luma-chroma demultiplexing algorithm for bayer demosaicking. IEEE Trans. Image Process. **20**(7), 1885–1894 (2010)
17. S Pei, I Tam, Effective color interpolation in ccd color filter arrays using signal correlation. IEEE Trans. Circuits Syst. Video Technol. **13**(6), 503–513 (2003)
18. X Lia, B Gunturk, L Zhang, Image demosaicing: a systematic survey. Proc. SPIE. **6822**, 68221J–68221J15 (2003)
19. K Chung, W Yang, W Yan, C Wang, Demosaicing of color filter array captured images using gradient edge detection masks and adaptive heterogeneity-projection. IEEE Trans. Image Process. **17**(12), 2356–2367 (2008)
20. L Zhang, X Wu, Color demosaicking via directional linear minimum mean square-error estimation. 14. **12**, 2167–2178 (2005)
21. K Chung, Y Chan, A low complexity color demosaicing algorithm based on integrated gradient. J. Electron. Imaging. **19**(2), 0211041–02110415 (2010)
22. A Rehman, L Shao, Classification-based de-mosaicing for digital cameras. Neurocomputing. **83**, 222–228 (2012)
23. SG Smith, Color image restoration with anti-alias. Patent US6842191, 2005
24. M Guarnera, G Messina, V Tommaselli, A Bruna, Directionally filter based demosaicing with integrated antialiasing. in *International Conference on Consumer Electronics, ICCE 2008. Digest of Technical Papers*, 1–2, (2008)
25. H Sheikh, M Sabir, A Bovik, A statistical evaluation of recent full reference image quality assessment algorithms. Image Process. IEEE. **15**, 3440–3451 (2006)
26. E Allen, S Triantaphillidou, R Jacobson, Image quality comparison between JPEG and JPEG2000. I. Psychophysical investigation. J. Imaging Sci. Technol. **51**, 548–258 (2007)
27. G Nyman, J Häkkinen, EM Koivisto, T Leisti, P Lindroos, O Orenius, T Virtanen, T Vuori, Evaluation of the visual performance of image processing pipes: information value of subjective image attributes. in *Proceedings of SPIE-IS&T Electronic Imaging*. (San Jose, California, vol. 7529, 752905-1–752905-10 2010)
28. T Leisti, J Radun, T Virtanen, R Halonen, G Nyman, Subjective experience of image quality: attributes, definitions and decision making of subjective image quality. in *Proceedings of SPIE-IS&T Electronic Imaging*. (San Jose, California, vol. 7242, 72420D-1–72420D-9, 2009)
29. J Radun, T Leisti, J Häkkinen, H Ojanen, J Olives, T Vuori, G Nyman, Content and quality: interpretation-based estimation of image quality. ACM Trans. Appl. Percept. **4** (2008)
30. V Kayargadde, J Martens, Perceptual characterization of images degraded by blur and noise: model. J. Opt. Soc. Am. A. **13**, 1178–1188 (1996)
31. GM Johnson, MD Fairchild, Sharpness rules. in *Proceedings of IS&T/SID 8th Color Imaging Conference*. (Scottsdale, 24–30, 2000)
32. M Cardaci, V Di Gesù, M Petrou, ME Tabacchi, A fuzzy approach to the evaluation of image complexity. Fuzzy Sets Syst. **160**(10), 1474–1484 (2009)
33. TN Pappas, RJ Safranek, J Chen. in *Handbook of Image and Video Processing*, Perceptual Criteria for Image Quality Evaluation. (San Diego: Academic Press, 939–959, 2005)
34. G Sharma, Digital Color Imaging Handbook. (CRC Press 2002)
35. L Thurstone, A law of comparative judgement. Psychol. Rev. **34**, 273–286 (1927)
36. P Longere, Z Xuemei, P Delahunt, D Brainard, Perceptual assessment of demosaicing algorithm performance. Proceedings of the IEEE. **90**(1), 123–132 (2002)

doi:10.1186/1687-6180-2012-123
Cite this article as: Gasparini *et al.*: A no-reference metric for demosaicing artifacts that fits psycho-visual experiments. *EURASIP Journal on Advances in Signal Processing* 2012 **2012**:123.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Immediate publication on acceptance
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com