J. Vis. Commun. Image R. 35 (2016) 236–247

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

# Evaluating color vision deficiency daltonization methods using a behavioral visual-search method ☆

Joschua Thomas Simon-Liedtke *, Ivar Farup

*The Norwegian Colour and Visual Computing Laboratory, Gjøvik University College, Pb. 191, 2802 Gjøvik, Norway*

ARTICLE INFO

*Article history:*
Received 20 August 2015
Accepted 21 December 2015
Available online 29 December 2015

*Keywords:*
Image processing
Daltonization
Color vision deficiency
Behaviorism
Visual-search
Response time
Accuracy
CVD simulation

ABSTRACT

Daltonization methods are used to automatically improve color images for color-deficient people. A comparison of different daltonization methods, however, is still left undone. We propose a visual-search method to evaluate daltonization methods by assessing behavioral performances of the attentional mechanism through the analysis of accuracy and response time data. Firstly, we show that the visual-search methodology can indeed be used to evaluate daltonization methods. Secondly, we argue that a combination of natural images and Ishihara images is needed to highlight differences between the daltonization methods. Our results indicate that the investigated daltonization methods can be ranked from highest to lowest as following: Firstly, the method proposed by Kotera; secondly, the method proposed by Fidaner et al.; thirdly, the method proposed by Huang et al.; and lastly the method proposed by Kuhn et al.



## 1. Introduction

A significant part of our society faces challenges in differentiating colors like for example the elderly or the color-deficient. Consequently, these groups face obstacles where colors are used as communication tool or information carrier – like for example in geographical or public transportation maps. Daltonization methods have been proposed to help the color-deficient by enhancing color contrast in natural images and/or information graphics.

However, a broad survey comparing and ranking daltonization methods, if they work and how good they perform, is still left undone. Also, there is not one standardized method to compare different daltonization methods.

We introduce the behavioral paradigm to evaluate daltonization methods in this paper. By analyzing accuracy and response time data from a visual-search experiment, we can compare and rank the individual methods. In our proposed experiment, we use images associated with visual-search tasks aimed to retrieve color information from the images. Observers are asked to complete the tasks by pressing answer keys on a keyboard, after which response time and correctness of the answer are recorded. We discuss in this paper, (i) if individual daltonization methods show a statistically significant difference in accuracy and/or response time; (ii) if and how these differences can be used to rank daltonization methods; (iii) what types of images work best for the proposed methodology, including natural and Ishihara images.

## 2. Background

The human visual system (HVS) of a normal-sighted person can distinguish between millions of different colors under certain viewing conditions. Cones, which are photoreceptors on the retina of the human eye, contain pigments that are sensitive to light of different wavelengths: So-called S-, M, and L- cones filter light of short, medium and long wavelengths [1, Chapter 2,2, Chapter 6]. This is also known as *trichromacy* [1, Chapter 5,2, Chapter 6].

For people with color vision deficiency (CVD), however, either the sensitivity of the cone pigments is slightly shifted in comparison to normal-sighted people (*anomalous trichromacy*), or certain cone pigments are missing (*dichromacy*) [3, Chapter 4,2, Chapter 6]. This leads to a decreased ability to differentiate certain colors, and/or not being able to perceive certain colors at all. Surveys suggest that about 8% of the male population are color-deficient [4]. CVDs along the red–green axis are most common [3, Chapter 4,2, Chapter 6]: In deutan anomalous trichromats, the sensitivity of the M-cones is shifted; in deuteranopes, the M-cones are missing completely. The respective terms for deviations in L-cones are protan anomalous trichromats and protanopes. CVD tests are used to

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
*E-mail address:* joschua.simonliedtke@hig.no (J.T. Simon-Liedtke).

http://dx.doi.org/10.1016/j.jvcir.2015.12.014
1047-3203/

identify CVDs, and/or categorize these CVDs according to type and severeness. Commonly used CVD tests include, for example, the Ishihara test [5], the Hardy-Rand-Rittler (HRR) test [6], the Farnsworth D15 test [7], the Lanthony D15 desaturated test [8] or the anomaloscope [9].

Daltonization methods are automatic image enhancement methods that improve visual experience for color-deficient people [10–16]. This is often done by increasing color contrast between confusion colors, i.e. colors that look different for the normal-sighted but (close-to) identical for the color-deficient. Anagnostopoulos et al. [10] cited a daltonization method proposed by Fidaner, Lin and Ozguven that translates color contrast of confusion colors into lightness contrast. This is done by simulating the color of each pixel in the original image $I(x,y)_0$, as a 3-by-1 vector, using a CVD simulation method in sRGB resulting in the simulated color $I(x,y)_P$ for protanopes and in the simulated color $I(x,y)_D$ for deuteranopes respectively. Secondly, the error difference $I(x,y)_E$ is computed by subtracting the simulated color from the original color. Finally, the error $I(x,y)_E$ is added to the original color $I(x,y)_0$ in order to obtain the daltonized color $I(x,y)_{dalt}$ by distributing it according to the matrix below that is optimized for protanopes:

$$I(x,y)_{dalt} = I(x,y)_0 + M \cdot I(x,y)_E \tag{1}$$

$$\text{where } M = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.7 & 1.0 & 0.0 \\ 0.7 & 0.0 & 1.0 \end{bmatrix} \tag{2}$$

Kotera's daltonization method [16] shifts the hue of confusion colors in order to become better visible after daltonization. It reintroduces lost spectral information for dichromats to the original image by a $\lambda$ hue shift that is determined through a cost function evaluating visibility for dichromats on the one hand, and visual gap between the original and daltonized image on the other hand. All calculations take place in pseudo-spectral subspaces of the LMS cone space, which are three-dimensional for the image steps representing normal-sighted and two-dimensional for the image steps representing dichromatic color vision. The basic idea of the following daltonization methods is that of maintaining, respectively optimizing the distance between so-called key colors. The daltonization method proposed by Huang et al. [17] uses a Gaussian Mixture Model (GMM) to group key colors in the original image, and optimizes the difference between the original key values and their simulations through the symmetric Kullback–Leibler (KL) divergence, a measure for dissimilarity between distributions. It produces non-deterministic image results, i.e the resulting daltonized images have different colors given the same input image and input variables for the algorithm (cf. Fig. 10). This is caused by the properties of the Gaussian Mixture Model (GMM), which creates different models for different runs of the GMM, as the author explains [18]. Kuhn et al. [19] proposed a daltonization method that uses mass-spring optimization on a quantized subset of key colors from the original image in CIELAB color space. By defining the inverse of the distance between original daltonized colors as mass, the system promises natural colors for the color-deficient. However, the resulting daltonized image remains in the reduced color space of dichromats.

In order to evaluate these daltonization methods, we have to somehow access the throughout subjective and internal sensation of color image perception. One way is through behaviorism, where psychological theories are founded on observable events [20, Page 14]. The goal of behaviorism is to make internal subjective processes empirically measurable by creating sensible stimuli that cause a measurable reaction in the observers [20, Page 14]. Methodologies used in behavioral science include visual-search tasks, in which a target stimuli is presented among so-called distractors that differ in either shape, color, size and/or texture, etc. [21]. The performance of the attentional mechanism is measured by recording and evaluating the *response time* of the observations, and whether or not the observer has *answered correctly*. Behavioral experiments can benefit the field of color imaging. In a previous paper [22], for example, we show how a task-based visual-search method can be used to compare daltonization methods on information graphics and geographic maps.

Daltonization methods should also be evaluated in the light of universal design, which is defined as design that enables a product's usage for the greatest extent of user groups without any further adjustment for any of the specific groups [23]. One principle of universal design is equitable use, i.e. identical or equivalent functionalities for both normal-sighted and color-deficient observers [24].

## 3. Method and implementation

In a preliminary study [25], we presented a method to access behavioral performance of the human visual system (HVS) connected to the attentional mechanism. Observers are asked to complete visual-search tasks based on retrieving color information from images. For the data analysis, response times (RTs) and accuracies are measured. In this paper, we show that measuring the behavioral performance of the HVS for different daltonization methods can be used to evaluate, compare and rank them.

The experiment is based on a number of images that are organized in different sets, motives, variants, and versions as represented by Fig. 1.

1. Images are grouped into sets according to similar information content like for example images of wrestlers, Ishihara plates, colored powders (cf. Fig. 2a–c), etc. Each set is associated with a common visual-search task, whose goal is to retrieve color information from key elements in the image. The task takes form of a statement like "The wrestlers have jerseys of different color hues" or "There is a number in the [Ishihara] image", etc. Each task exists in two formulations: The first of which addresses color differences ("The wrestlers have jerseys of different color hue"), whereas the second addresses color similarities ("The wrestlers have jerseys of the same color hue").
2. Each set contains at least one and up to four different motives with similar image content (cf. Fig. 2d–f) in order to span as many as possible of the image quality classification attributes proposed by Pedersen et al. [26] with minor adjustments for the color-deficient: predominant color[1] hue/s (red, yellow, green, cyan, blue, magenta and/or multicolored), protan/deutan/tritan/normal color contrast, overall lightness and saturation, memory colors like skin color (African, Asian and/or Caucasian), sky-blue and/or grass-green, uniform colored/neutral areas, color transitions, fine details, and busyness. I.e. there are images that are neutral, colorful, bright, dark, blurry and/or sharp, etc. Table 1 shows an overview over all 13 sets, their associated tasks and the 22 motives. Sets 1–10 contain natural images, whereas sets 11–13 contain Ishihara plates.
3. Different motive variants have been created in order to minimize confounding bias. Key colors, i.e. colors of key objects and/or areas carrying relevant color information with respect to the associated task, have been altered manually for each variant in Adobe Photoshop CS6. In general, only the hue of these key colors has been changed, while preserving both chroma and lightness. Each motive exists in up to three different

[1] For interpretation of color in Figs. 2, 3 and 10, the reader is referred to the web version of this article

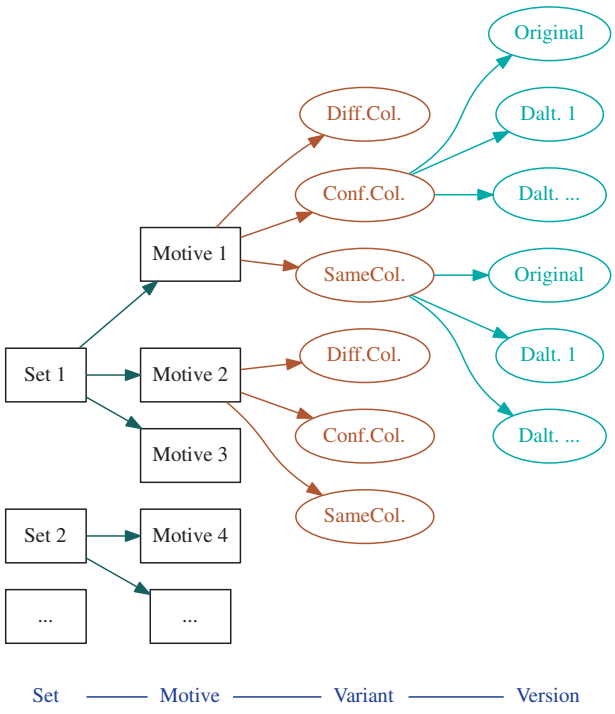



**Fig. 1.** Image data base structure visualized as tree.

variants representing different degrees of difficulty for color-deficient and normal-sighted observers: (a) A variant with different colors (Diff.Col.) (cf. Fig. 3a), in which the chosen key colors are easy to differentiate for both normal-sighted and color-deficient observers equally. (b) A variant, in which the key colors are changed to confusion colors (Conf.Col.) (cf. Fig. 3b) that are easy to differentiate for normal-sighted but difficult to differentiate for color-deficient observers. Confusion colors have been identified and verified with the help of the color-deficiency simulation methods proposed by Brettel et al. [15] and Kotera [16]. Confusion colors have been adjusted for different CVDs, i.e. along the axis blue–green–gray–red for protan, along green–gray–red–purple for deutan, and along violet–gray–yellow–green for tritan CVDs [3, Chapter 5]. And (c) a variant with same colors (SameCol.) (cf. Fig. 3c), in which the key colors are identical for both normal-sighted and color-deficient observers. Each variant has a unique and unambiguous answer with respect to the task associated with the superordinate set, i.e. the answer is either true or false. Some motives have only a Conf.Col. and SameCol. variant depending on the statement. This concerns mostly images with statements connected to color naming – "There is green powder in the image", and the Ishihara plates – "There is a number in the image."

4. The daltonized image versions for each of the Conf.Col. and SameCol. variants serve as target stimuli in the experimentation. We chose the methods proposed by Fidaner et al. [10], by Kotera [16], by Kuhn et al. [19], and by Huang et al. [17] (cf. Fig. 3d–g). Lastly, we included a control color-to-gray method, which can be thought of as "worst case" scenario (cf. Fig. 3h, called the Dummy method). The original version of each variant without daltonization is also defined as one version of the image.

Once the image data base is defined, it can be used for the Visual-Search Daltonization Evaluation Method (ViSDEM) (cf. Fig. 4): (i) The statement related to a random set is presented on the screen. (ii) After a fixation screen is shown for 500 ms, the first target image of the set is presented. The observer is asked to answer as quickly and as correctly as possible, whether he/she agrees or disagrees to the statement for the target image. (iii) The program records the response time (RT) and whether or not the observer has answered correctly to the task. The target images presented in the experimentation are randomized by motive, variant, version, and task formulation in order to avoid multiple biases.

We expect certain observations for the behavioral data:

1. If the presented experimental design works correctly, we expect differences in behavioral performance between the original non-daltonized variants for normal-sighted and color-deficient observers. (i) We expect similar response times and accuracies for the Diff.Col. variants for both color-deficient and normal-sighted observers. (ii) In contrast, we expect higher response times and lower accuracies for the Conf.Col. variant of color-deficient than of normal-sighted observers. This is because color-deficient observers are supposed to make more mistakes or need more time in distinguishing the colors in the original Conf.Col. variants.
2. Optimal daltonization methods should lead to improved behavioral performance of color-deficient observers for the Conf.Col. variant. More precisely, the higher the accuracies and the lower the response times the better the daltonization method. Moreover, a good daltonization method does not decrease behavioral performance of normal-sighted people in the light of universal design.

The statistical methods for the analysis of the visual-search data were identical to the ones used in a previous article [27]: Namely, the computation of the accuracy confidence intervals with the Wilson interval score [28], the analysis of the accuracy data with the unpaired $\chi^2$ test and the paired student-$t$ test [29, Chapter 8], the analysis of the response time data with the Mood's median test [30, Pages 394–399]. Furthermore, we plotted the results of the response time (RT) data as a box plot, in which the median RT is indicated by a red line, and the confidence interval of the median is represented by a notch computed using a Gaussian-based asymptotic approximation [31]. The visual-search method was implemented with PsychoPy2 [32], and multiple Python libraries were used for the statistical data analysis, namely the NumPy, the Matplotlib, the SciPy, and the Pandas libraries [33–36].

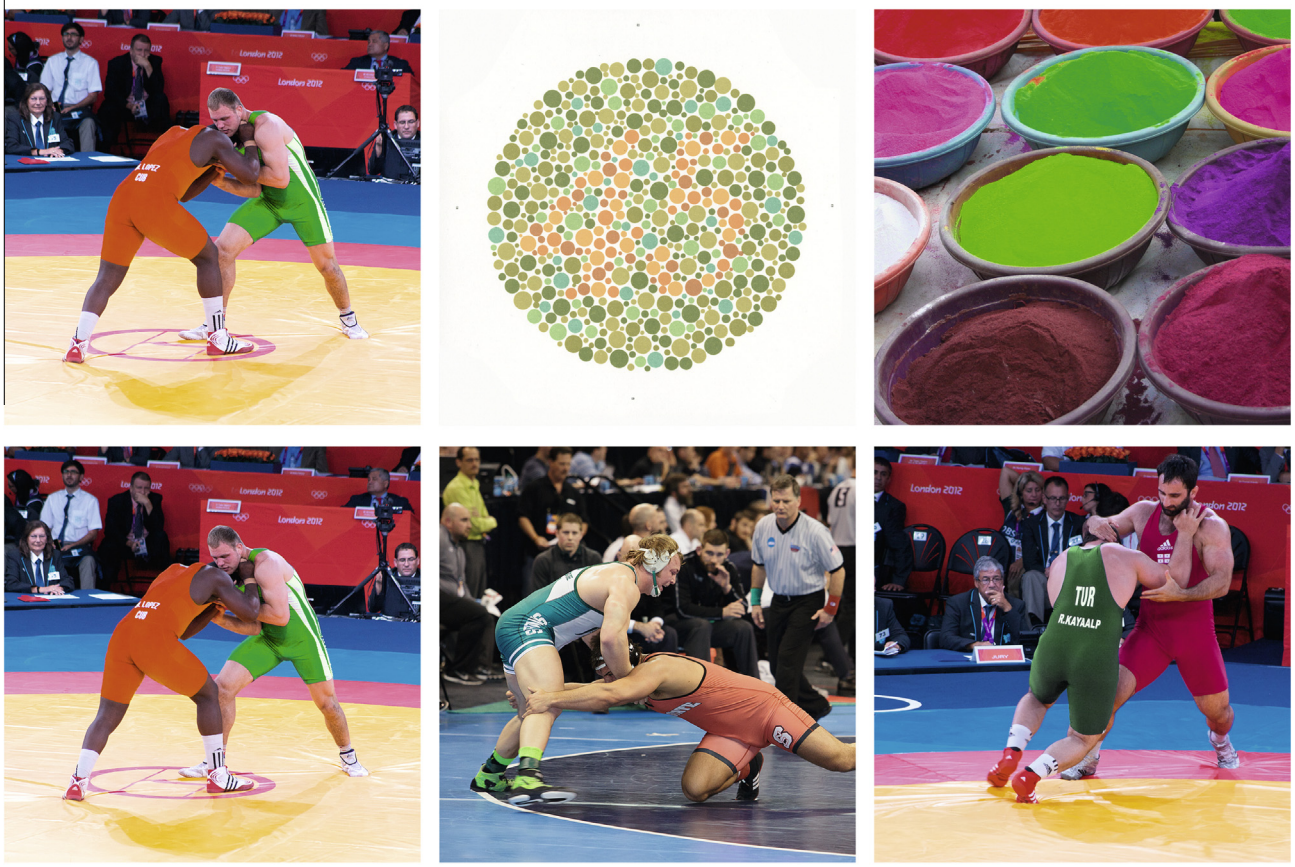
Every observer was asked to take each of four different CVD tests: the Ishihara test, the HRR test, the Farnsworth D15 test and the Lanthony desaturated D15 test. Especially, the HRR test was used to investigate type and severeness of the CVD, and the Farnsworth D15 test and Lanthony D15 test was used to confirm the results of the HRR test. A separation between anomalous trichromats and dichromats, however, was not possible with the applied CVD tests, which would require an anomaloscope. There were 25 observers in total, 10 with normal color vision, and 15 with some sort of deutan CVD. According to the HRR test, 9 observers showed signs of strong, 4 observers showed signs of medium, and 2 observers showed signs of mild deutan CVD. We focused on deutan color deficient observers in this paper because they make up about 75% of all color-deficient people [3, Chapter 3].

We conducted the experiments on two PCs running Windows 7 with identical setups and calibrations. They were calibrated with Eye-One Match Pro to fit a medium white, a gamma of 2.2 and an illuminance of 120 lux. The surrounding lights were D50-like fluorescent lights dimmed to ca. 200 lux (resulting in a color temperature of 4230 K) for the CVD tests, and dimmed to ca. 30 lux (4411 K) for the experiment. The luminance of the table

(a) Set: The wrestlers have jerseys of different color hue. Correct answer: Agree.

(b) Set: There is no number in the image. Correct answer: Disagree.

(c) Set: Some of the powder in the boxes is green. Correct answer: Agree.




(d) Motive one of the wrestler set.

(e) Motive two of the wrestler set.

(f) Motive three of the wrestler set.

**Fig. 2.** Top: Examples of different sets in the image data base, the corresponding visual-search task, and the corresponding correct answer. Bottom: Examples of different motives within one set of the image data base.

for the CVD tests was about 38 $\frac{\text{cd}}{\text{m}^2}$ measured with a Minolta CS 100 spectroradiometer under a $45^\circ \setminus 0^\circ$ viewing angle.

## 4. Results and analysis

### 4.1. Analysis of the variants

The variant accuracies for deutan color-deficient observers can be seen in Fig. 5a: The Diff.Col. variant with 0.91 (from 174 observations), the Conf.Col. variant with 0.74 (from 256 observations), and the SameCol. variant with 0.87 (from 250 observations). Likewise, the variant accuracies for normal-sighted observers can be seen in Fig. 5b: The Diff.Col. and the Conf.Col. variants with both 0.99 (from 104 and 158 observations), and the SameCol. variant with 0.93 (from 152 observations). The statistical analysis of the variant data supports the general validity of the experimental design (cf. Table 2a and b): (i) The accuracy of the Conf.Col. variants is statistically significantly lower than the accuracy of the Diff.Col. variants for deutan color-deficient observers as expected. In contrast, there is no statistically significant difference between both variants for normal-sighted observers. This observation supports the assumption that the correct confusion colors have been chosen for the Conf.Col. variants. (ii) The accuracy of the Diff.Col. variant for deutan color-deficient observers is statistically significantly lower than the accuracy of the Diff.Col. variant for normal-sighted observers as expected. This indicates that the HVS' attentional mechanism of the color-deficient has a slight empirical disadvantage compared to the normal-sighted as discussed in a previous paper [37]. (iii) For deutan color-deficient observers, the accuracy of the SameCol. variant is significantly different from the Conf.Col. variant, but not from the Diff.Col. variant. In contrast, the accuracy of the SameCol. variant is statistically significantly lower than both the Diff.Col. and the Conf.Col. variants for normal-sighted observers. Some normal-sighted observers might have interpreted color lightness differences of the SameCol. variants as color hue differences.

As is common for the analysis of response time data, only the RTs of correct observations have been considered [38–40]. Only the median RT of the SameCol. variant is statistically significant higher than the Conf.Col. variants for deutan color-deficient observers (cf. Fig. 5c and Table 2c). Some color-deficient observers might consciously use more time on images that appear similar. The RTs for normal-sighted observers do not give any meaningful feedback (cf. Fig. 5d and Table 2d).

**Table 1**
Overview of the sets, the associated statements, and descriptions about the motives in the sets. Also, the table contains information whether the confusion colors are difficult for protan (Prot.), deutan (Deut.) or tritan (Trit.) color-deficient people.

| Set | Motives | Confusion colors | Statement |
|---|---|---|---|
| 1 | Wrestlers in colored jerseys | Prot./ Deut. | The wrestlers have jerseys of different color hue |
| 2 | Berries and fruits in front of foliage | Prot./ Deut. | The berries/fruits have a different color hue as the leaves in the background |
| 3 | Two people with colored tops | Prot./ Deut. | Both people wear tops of different color hue |
| 4 | Couple with roses | Prot./ Deut. | At least one rose in the image has a different color hue |
| 5 | Colored chalk roses on concrete | Prot./ Deut. | The chalk blossom has a different color hue from its painted stem |
| 6 | Colored chalk hearts on concrete | Prot./ Deut. | Some of the chalk hearts have a different color hue |
| 7 | Colored powders | Prot./ Deut. | Some of the powder in the boxes is green |
| 8 | Paprikas | Prot./ Deut. | At least one paprika has a different color hue than the other |
| 9 | Feather in front of foliage on the ground | Trit. | The feather has a different color hue than the leaves in the background |
| 10 | Bird and frog in front of foliage and grass | Trit. | The animal has a different color hue than the plants in the background |
| 11 | Ishihara plates#13 and #20 | Prot./ Deut. | There is a number in the image |
| 12 | Ishihara plates#17 and #21 | Prot./ Deut. | There is a number in the image |
| 13 | Ishihara plates#22, #23, #24 and #25 | Prot./ Deut. | There is a number in the image |

### 4.2. Analysis of individual sets

A detailed analysis of the Conf.Col. variants in each set reveals how evenly the sets measure behavioral performance. The accuracies for deutan color-deficient observers show huge variances between the different sets (cf. Fig. 6a). Most differences are related to the distinction between natural, sets 1–10, and Ishihara sets, sets 11–13. The natural sets have a statistically significant higher accuracy than the Ishihara sets (cf. Fig. 7). It is difficult to create natural images with confusion colors only, while preserving the overall color naturalness of key objects in the images, whereas Ishihara images are specifically designed for being unreadable for the color-deficient. In contrast, accuracies of normal-sighted observers are very homogeneous throughout all sets as expected (cf. Fig. 6b).

### 4.3. Analysis of individual daltonization methods for deutan color-deficient observers

Only the data from the Conf.Col. variants are relevant for the analysis of the individual daltonization methods since these variants contain confusion colors that manifest the biggest problems for color-deficient observers. The accuracies of the individual methods for deutan color-deficient observers can be ranked from highest to lowest (cf. Fig. 8a): The Kotera method with 0.89, the Fidaner method with 0.82, the Huang method 0.77, the original version with 0.74, the Kuhn method with 0.73 and the Dummy method with 0.10. All accuracies are obtained from 256 individual observations. The statistical analysis allows segmentation in four distinct groups (cf. Table 3a): (i) The Kotera method has the highest accuracy compared to all other methods including the original.

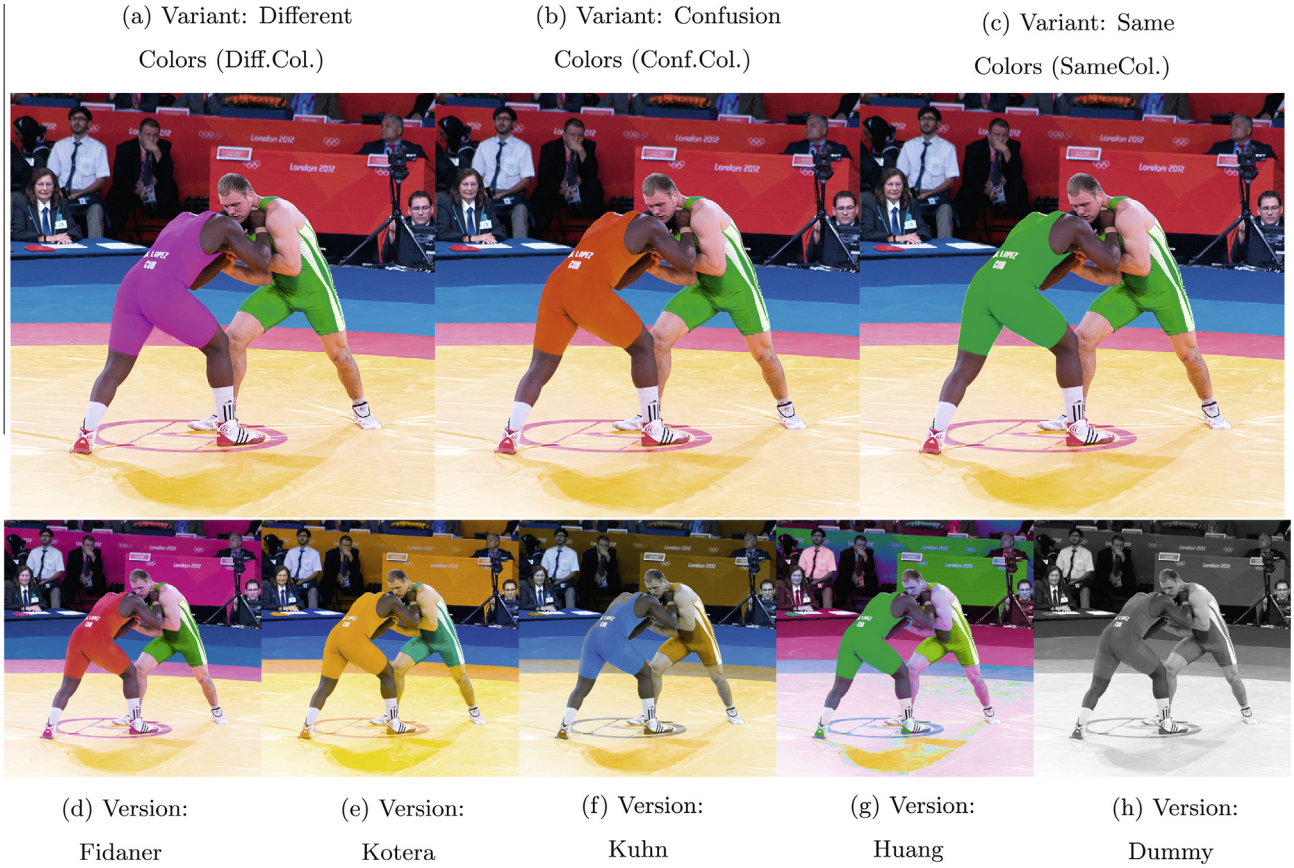



**Fig. 3.** Example images of different variants of a motive (top) and versions of the Conf.Col. variant (bottom).

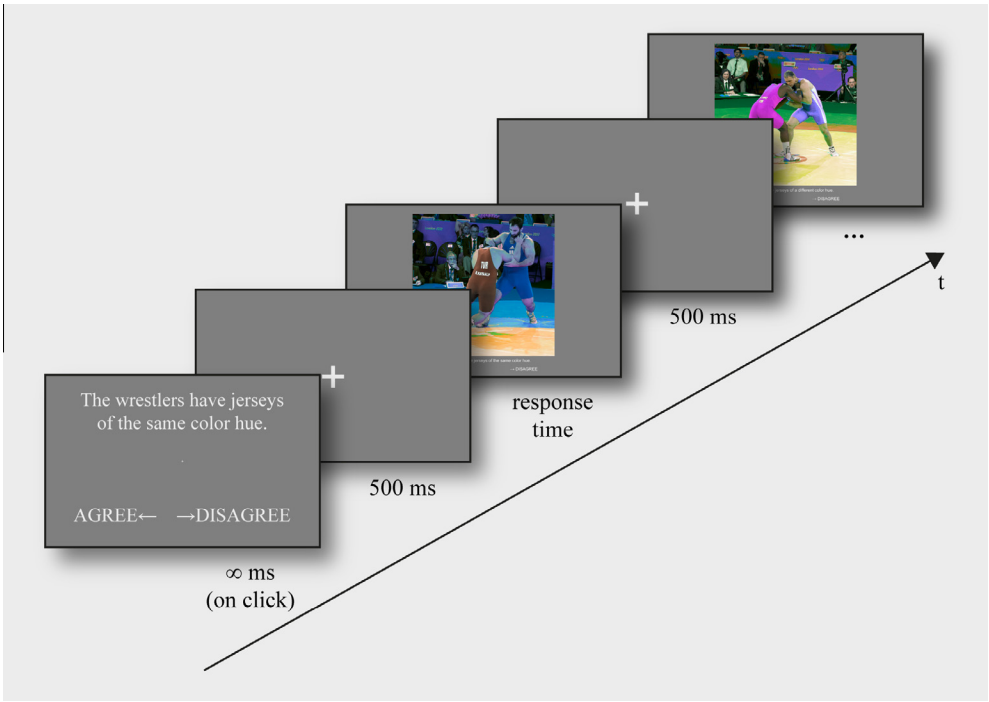
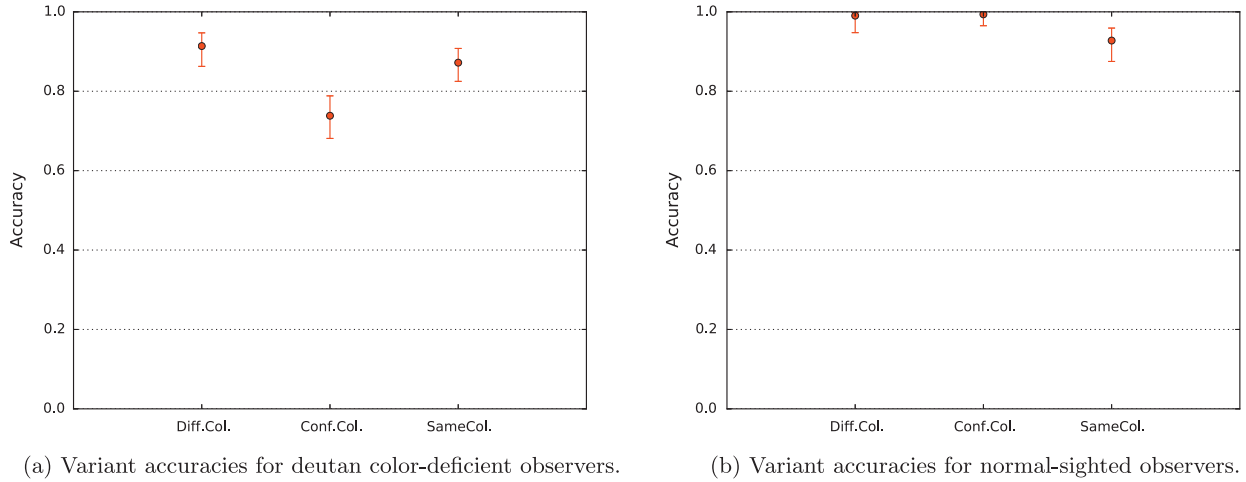


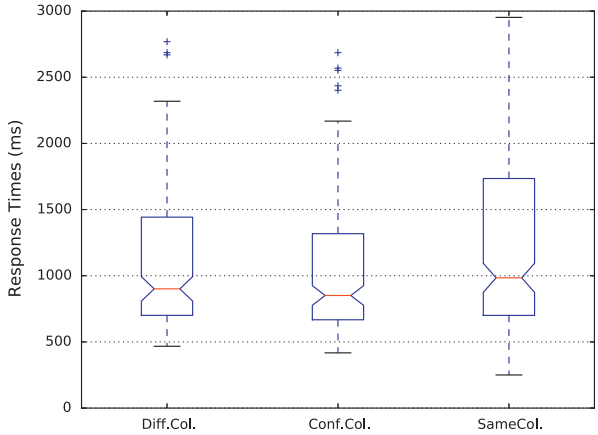

**Fig. 4.** The workflow of the proposed method: At first, the statement related to the set and the answer key options are shown. Secondly, the target image is presented. The objective is to answer as quickly and as accurately as possible. Thirdly, the program moves on to the next target image, after response time and correctness are recorded.



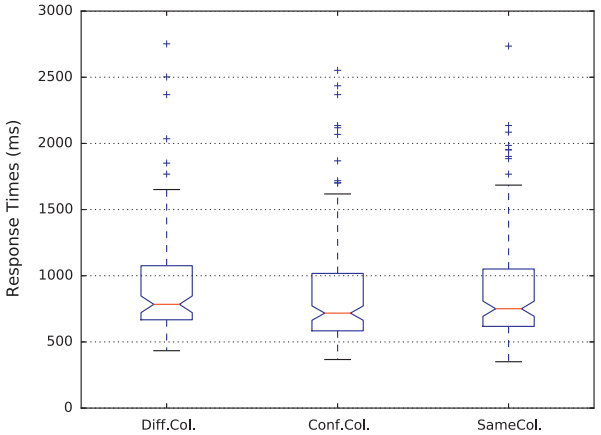

(a) Variant accuracies for deutan color-deficient observers.




(b) Variant accuracies for normal-sighted observers.




(c) Variant response times for deutan color-deficient observers.




(d) Variant response times for normal-sighted observers.

**Fig. 5.** Accuracy (top) and response time data (bottom) grouped by variants. All original versions of each variant are collapsed.

**Table 2**
Statistical analysis of the accuracy data (2a and b) and response time data (2c and d) between the individual variants. All original versions of each variant are collapsed. Statistically significant values are emphasized in bold.

| | Conf.Col. | SameCol. |
|---|---|---|
| *(a) p-Values of the $\chi^2$ test for deutan color-deficient observers* | | |
| Diff.Col. | $\mathbf{5.4 \times 10^{-6}}$ | 0.18 |
| Conf.Col. | x | $\mathbf{1.5 \times 10^{-4}}$ |
| *(b) p-Values of the $\chi^2$ test for normal-sighted observers* | | |
| Diff.Col. | 0.76 | **0.02** |
| Conf.Col. | x | $\mathbf{2.6 \times 10^{-3}}$ |
| *(c) p-Values of the Mood's median test for deutan color-deficient observers* | | |
| Diff.Col. | 0.45 | 0.27 |
| Conf.Col. | x | **0.04** |
| *(d) p-Values of the Mood's median test for normal-sighted observers* | | |
| Diff.Col. | 0.10 | 0.36 |
| Conf.Col. | x | 0.42 |

(ii) The second highest accuracy can be observed for the Fidaner version, which is higher than all other daltonization methods and the original except for the Huang method. (iii) In contrast, daltonization with the Huang and Kuhn methods does not lead to increased accuracies compared to the original. And (iv) the Dummy method has the lowest accuracy as expected, which supports the general validity of our experimental design.

Considering Section 4.2, it makes sense to analyze the results in groups of natural (cf. Fig. 8c) and Ishihara images (cf. Fig. 8d) as well. The overall ranking for "natural images only" is very similar to the ranking obtained for both image groups combined, at the same time as the statistical significant differences become less clear (cf. Table 3b). The Kotera method has still the highest accuracy, and the Huang and the Kuhn methods do not increase accuracy in comparison to the original. However, the Huang and the Kuhn methods actually improve behavioral performances as compared to the original "Ishihara images only" (cf. Table 3c). This is of no surprise because many daltonization methods focus on the improvement for especially these types of images.

The response time data is somewhat insignificant (cf. Fig. 8b). Only the median RT for the Kotera method is statistically significantly higher than the original and the Fidaner method (cf. Table 3d). However, we expected a lower median RT for the Kotera method because of its higher accuracy. The Kotera method often creates unnatural colors in natural images, which might lead to higher RTs as pointed out by Bramão et al. [39]. But in general, no further conclusion for the evaluation of daltonization methods can be drawn from the response time data.

### 4.4. Analysis of individual daltonization methods for normal-sighted observers

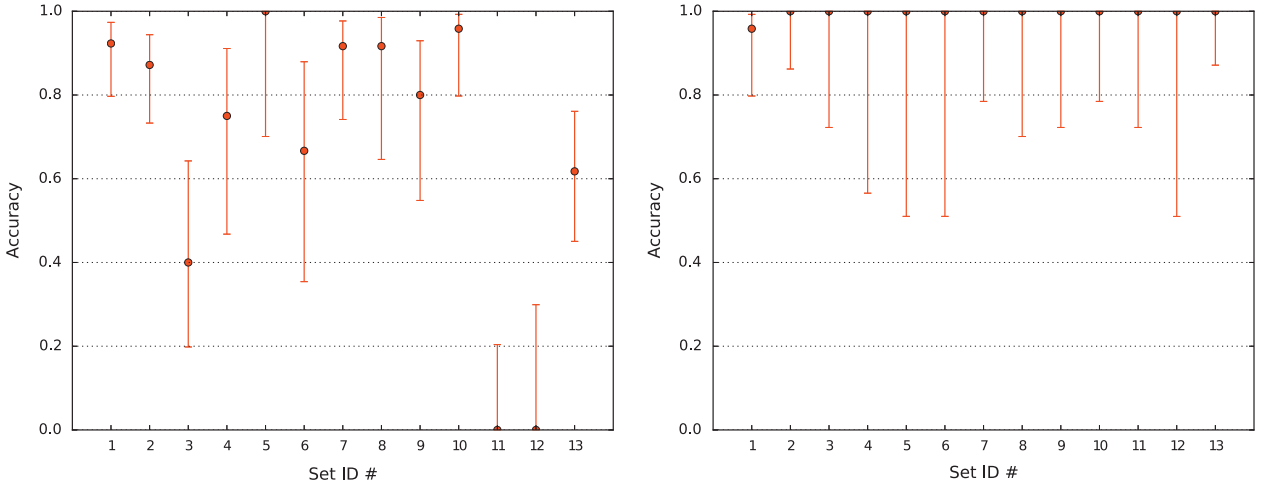
The ranking of the daltonization methods according to accuracies for normal-sighted observers is somewhat different (cf. Fig. 9a): The original and the Fidaner method with 0.99 each, the Kotera method with 0.97, the Huang method with 0.88, the Kuhn method with 0.77, and finally the Dummy method with 0.03. All accuracies have been computed from 158 observations. The statistical analysis reveals four distinct groups (cf. Table 4a): (i) The Fidaner and Kotera methods have close to optimal accuracies. However, since the accuracy for the original is already high, it would be virtually impossible to increase accuracy. (ii) The Huang method decreases accuracy in comparison to both the original and the Fidaner and Kotera methods. (iii) Again, the Kuhn method has a lower accuracy than any other daltonization method and the original. Both the Huang and Kuhn methods, which intend to enhance images for the color-deficient, seem to have the side effect of decreasing behavioral performance for the normal-sighted. (iv) The Dummy method has the lowest accuracy, which is in accordance with our prediction and supports the general validity of our experimental setup.

Analyzing natural and Ishihara images individually gives only one additional insight (cf. Fig. 9c and d): The Huang method performs slightly better on "Ishihara images only" than on "natural images only", whereas the Kuhn method still decreases behavioral performance for both image types (cf. Table 4b and c).

Again, the response time data does not give any relevant feedback (cf. Fig. 9b and Table 4d) as observed for deutan color-deficient observers before.

## 5. Discussion

In the following discussion, we focus on the analysis of "natural and Ishihara images combined". We argue that relying on both image types is important. First of all, color distributions are very different for Ishihara and natural images: In the former, uniform color areas are separated by clear edges; in the latter, color transitions, color gradients and noise are more present. Consequently, some daltonization methods work best for natural images, whereas




(a) Accuracies for deutan color-deficient observer.

(b) Accuracies for normal-sighted observer.

**Fig. 6.** Accuracy data for the Conf.Col. variants of the original motives grouped by set IDs.

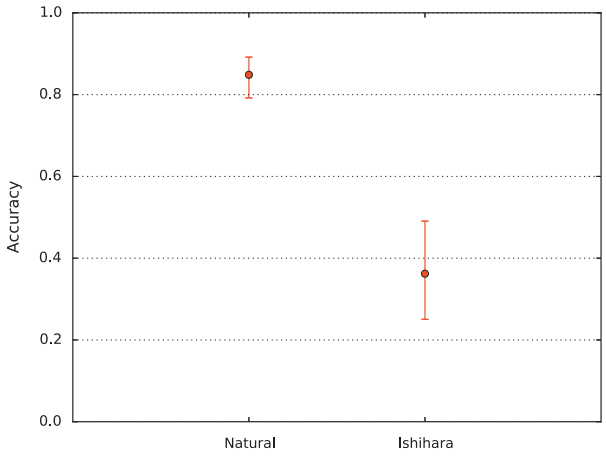
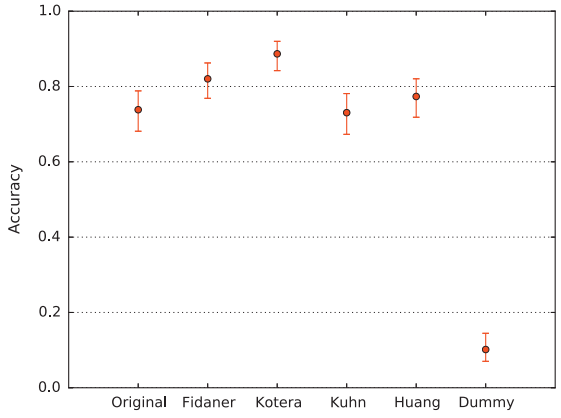


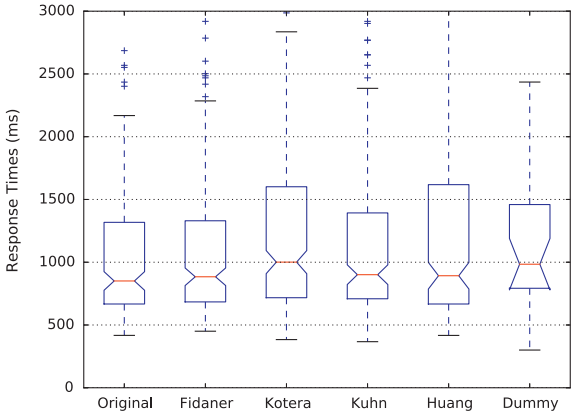

**Fig. 7.** Accuracies for deutan color-deficient observer.

others work best on the Ishihara plates. A combination of both image types for the analysis allows the coverage of as many different types of images as possible. This is valid since the analysis from the previous section confirms that the overall ranking order of the daltonization methods is very similar for both natural and Ishihara images only. At the same time, we chose more natural images than Ishihara images because natural images are more present in daily life. This allows us to emulate behavioral performance in day-to-day situations. Finally, the accuracy differences for natural images are more subtle because natural images do not have as clearly defined confusion colors as the Ishihara images. Thus, a higher number of data points is required in order to obtain statistical significant differences. Daltonization methods provide balanced color image enhancement for the color-deficient being cost-efficient, easy to implement and intuitive to use as discussed in a previous paper [41]. Also, a daltonization method that performs well in the light of universal design *increases* behavioral performance for the color-deficient, at the same time as it does not *decrease* behavioral performance for the normal-sighted.

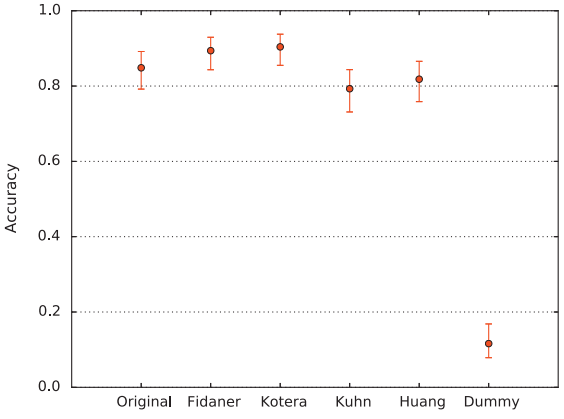
According to these premises, the chosen daltonization methods are ranked as following: (1) the Kotera method, (2) the Fidaner method, (3) the Huang method, and (4) the Kuhn method. The ranking reveals good and bad daltonization strategies: (i) The Kotera method can produce somewhat unnatural colors: Leaves turning blue, a yellow tint in images that have areas of uniform colors/gray (cf. Fig. 3e), etc. However, this did not seem to influence behavioral performance negatively. (ii) The Fidaner method is originally optimized for protanopes, but it turns out to be quite effective for deutan color-deficient observers as well. Indeed, protan and deutan confusion colors are very similar [15]. Also, the Fidaner method uses protan and deutan CVD simulations respectively in order to compute the error image. The algorithm could be



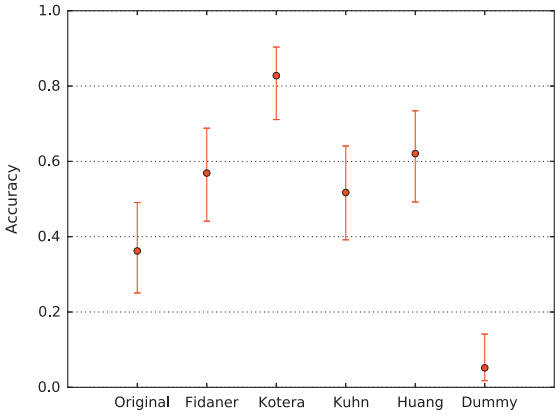

(a) Accuracies of "natural and Ishihara images combined".




(b) Response times of "natural and Ishihara images combined".




(c) Accuracies of "natural images only".

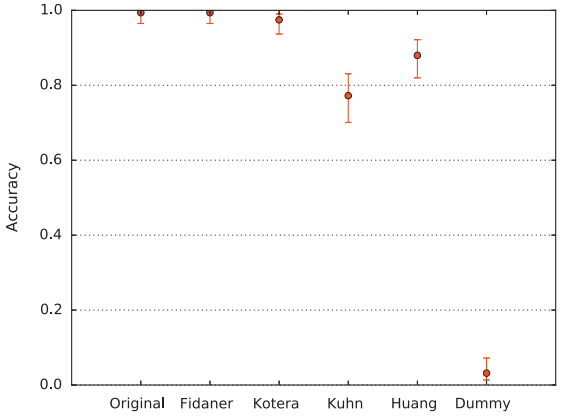



(d) Accuracies of "Ishihara images only".

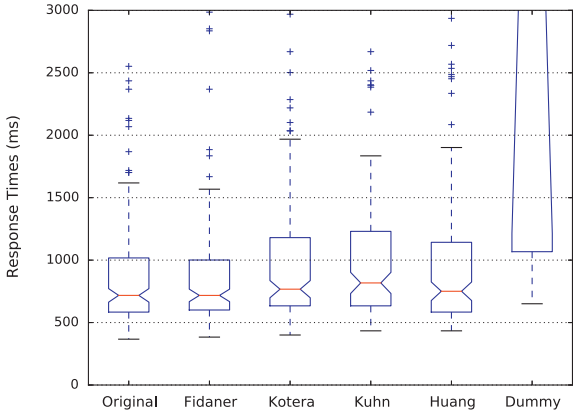
**Fig. 8.** Accuracy (cf. a, c and d) and response time data (cf. b) grouped by daltonization methods for deutan color-deficient observers. All Conf.Col. variants of each daltonization method are collapsed.

**Table 3**
Statistical analysis of the accuracy (3a–c) and response time data (3d) between the individual daltonization methods for deutan color-deficient observers. All Conf.Col. variants of each daltonization method are collapsed. Statistically significant values are highlighted in bold.

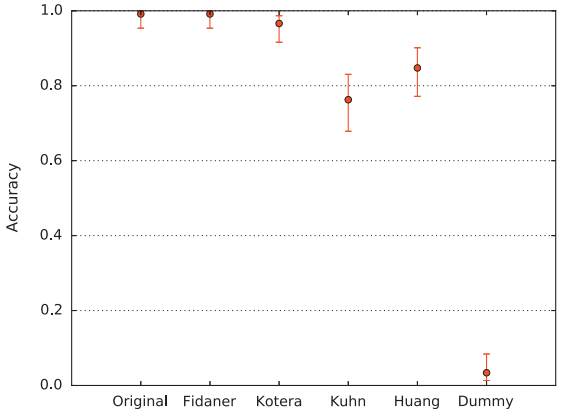
| | Fidaner | Kotera | Kuhn | Huang | Dummy |
|---|---|---|---|---|---|
| *(a) Paired student-t test p-values of "natural and Ishihara images combined"* | | | | | |
| Original | $\mathbf{1.1 \times 10^{-3}}$ | $\mathbf{9.1 \times 10^{-7}}$ | 0.59 | 0.41 | $\mathbf{4.3 \times 10^{-43}}$ |
| Fidaner | x | **0.01** | $\mathbf{7.9 \times 10^{-4}}$ | **0.03** | $\mathbf{1.1 \times 10^{-54}}$ |
| Kotera | x | x | $\mathbf{6.8 \times 10^{-8}}$ | $\mathbf{2.6 \times 10^{-5}}$ | $\mathbf{6.7 \times 10^{-64}}$ |
| Kuhn | x | x | x | 0.16 | $\mathbf{6.0 \times 10^{-42}}$ |
| Huang | x | x | x | x | $\mathbf{7.9 \times 10^{-49}}$ |
| *(b) Paired student-t test p-values of "natural images only"* | | | | | |
| Original | 0.06 | **0.04** | 0.10 | 0.34 | $\mathbf{2.3 \times 10^{-43}}$ |
| Fidaner | x | 0.66 | $\mathbf{1.9 \times 10^{-3}}$ | $\mathbf{8.7 \times 10^{-3}}$ | $\mathbf{1.1 \times 10^{-49}}$ |
| Kotera | x | x | $\mathbf{3.0 \times 10^{-4}}$ | $\mathbf{4.9 \times 10^{-3}}$ | $\mathbf{4.6 \times 10^{-52}}$ |
| Kuhn | x | x | x | 0.46 | $\mathbf{2.4 \times 10^{-38}}$ |
| Huang | x | x | x | x | $\mathbf{2.0 \times 10^{-43}}$ |
| *(c) Paired student-t test p-values of "Ishihara images only"* | | | | | |
| Original | $\mathbf{3.5 \times 10^{-3}}$ | $\mathbf{2.2 \times 10^{-8}}$ | **0.03** | $\mathbf{2.7 \times 10^{-4}}$ | $\mathbf{6.4 \times 10^{-4}}$ |
| Fidaner | x | $\mathbf{4.4 \times 10^{-4}}$ | 0.21 | 0.77 | $\mathbf{8.5 \times 10^{-8}}$ |
| Kotera | x | x | $\mathbf{7.7 \times 10^{-06}}$ | $\mathbf{2.7 \times 10^{-4}}$ | $\mathbf{3.2 \times 10^{-13}}$ |
| Kuhn | x | x | x | **0.02** | $\mathbf{6.2 \times 10^{-6}}$ |
| Huang | x | x | x | x | $\mathbf{1.2 \times 10^{-7}}$ |
| *(d) Mood's median test p-values of "natural and Ishihara images combined"* | | | | | |
| Original | 0.39 | $\mathbf{4.3 \times 10^{-3}}$ | 0.35 | 0.20 | 0.20 |
| Fidaner | x | **0.04** | 0.88 | 1.00 | 0.41 |
| Kotera | x | x | 0.37 | 0.35 | 0.69 |
| Kuhn | x | x | x | 0.88 | 0.66 |
| Huang | x | x | x | x | 0.68 |



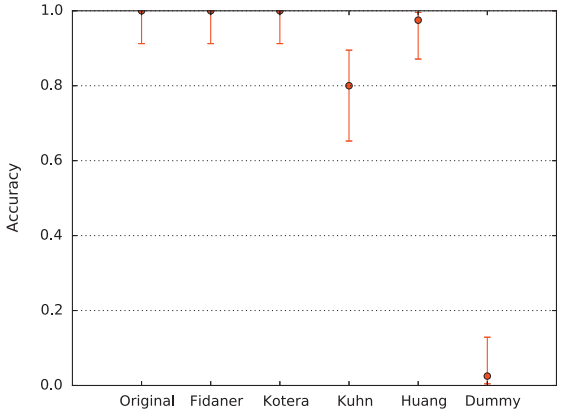

(a) Accuracies of "natural and Ishihara images combined".




(b) Response times of "natural and Ishihara images combined".




(c) Accuracies of "natural images only".




(d) Accuracies of "Ishihara images only".

**Fig. 9.** Accuracy (cf. a, c, and d and response time data (cf. b) grouped by daltonization methods for normal-sighted observers. All Conf.Col. variants of each daltonization method are collapsed.

**Table 4**
Statistical analysis of the accuracy (4a–c) and response time data (4d) between the individual daltonization methods for normal-sighted observers. All Conf.Col. variants of each daltonization method are collapsed. Statistically significant values are highlighted in bold.

| | Fidaner | Kotera | Kuhn | Huang | Dummy |
|---|---|---|---|---|---|
| *(a) Paired student-t test p-values of "natural and Ishihara images combined"* | | | | | |
| Original | 1.00 | 0.18 | $\mathbf{1.3 \times 10^{-9}}$ | $\mathbf{3.9 \times 10^{-5}}$ | $\mathbf{1.4 \times 10^{-111}}$ |
| Fidaner | x | 0.18 | $\mathbf{1.3 \times 10^{-9}}$ | $\mathbf{1.3 \times 10^{-5}}$ | $\mathbf{1.4 \times 10^{-111}}$ |
| Kotera | x | x | $\mathbf{1.4 \times 10^{-7}}$ | $\mathbf{1.6 \times 10^{-3}}$ | $\mathbf{6.4 \times 10^{-98}}$ |
| Kuhn | x | x | x | $\mathbf{4.8 \times 10^{-3}}$ | $\mathbf{7.9 \times 10^{-47}}$ |
| Huang | x | x | x | x | $\mathbf{6.9 \times 10^{-65}}$ |
| *(b) Student-t test p-values of "natural images only"* | | | | | |
| Original | 1.00 | 0.18 | $\mathbf{1.3 \times 10^{-7}}$ | $\mathbf{6.1 \times 10^{-5}}$ | $\mathbf{3.6 \times 10^{-82}}$ |
| Fidaner | x | 0.18 | $\mathbf{1.3 \times 10^{-7}}$ | $\mathbf{2.1 \times 10^{-5}}$ | $\mathbf{3.6 \times 10^{-82}}$ |
| Kotera | x | x | $\mathbf{1.1 \times 10^{-5}}$ | $\mathbf{2.5 \times 10^{-3}}$ | $\mathbf{3.2 \times 10^{-70}}$ |
| Kuhn | x | x | x | 0.07 | $\mathbf{6.0 \times 10^{-35}}$ |
| Huang | x | x | x | x | $\mathbf{1.7 \times 10^{-44}}$ |
| *(c) Student-t test p-values of "Ishihara images only"* | | | | | |
| Original | nan | nan | $\mathbf{3.3 \times 10^{-3}}$ | 0.32 | $\mathbf{7.8 \times 10^{-31}}$ |
| Fidaner | x | nan | $\mathbf{3.3 \times 10^{-3}}$ | 0.32 | $\mathbf{7.8 \times 10^{-31}}$ |
| Kotera | x | x | $\mathbf{3.3 \times 10^{-3}}$ | 0.32 | $\mathbf{7.8 \times 10^{-31}}$ |
| Kuhn | x | x | x | $\mathbf{6.4 \times 10^{-3}}$ | $\mathbf{4.0 \times 10^{-13}}$ |
| Huang | x | x | x | x | $\mathbf{2.9 \times 10^{-25}}$ |
| *(d) Mood's median test p-values of "natural and Ishihara images combined"* | | | | | |
| Original | 0.73 | 0.23 | **0.03** | 0.56 | 0.17 |
| Fidaner | x | 0.23 | 0.08 | 0.56 | 0.17 |
| Kotera | x | x | 0.33 | 0.45 | 0.17 |
| Kuhn | x | x | x | 0.29 | 0.17 |
| Huang | x | x | x | x | 0.17 |

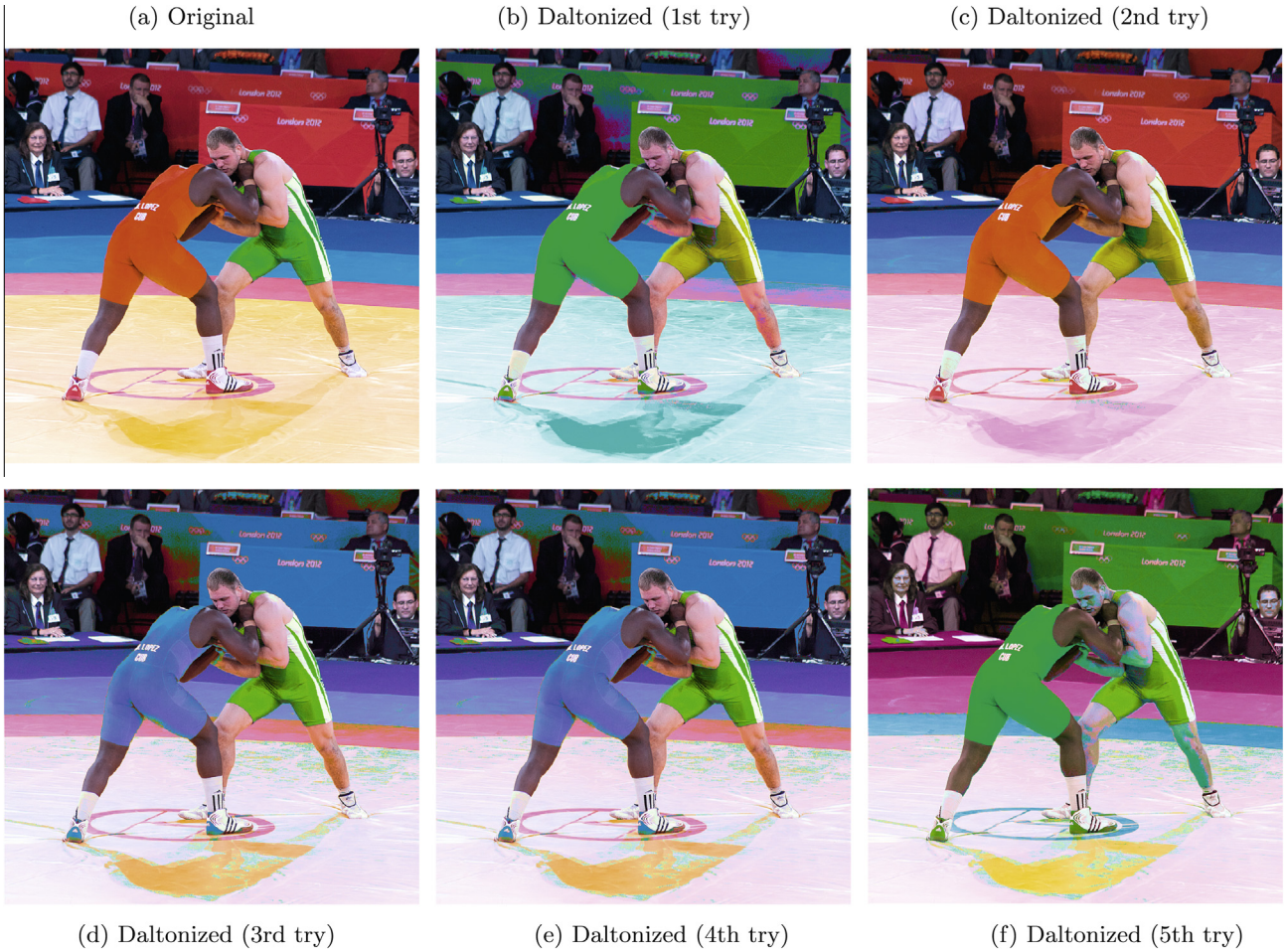



**Fig. 10.** Examples of the non-deterministic aspect of deuteranopia daltonization using the Huang method.

improved for deutan color-deficient observers by using a different error distribution matrix $M$ in Eq. (1):

$$M = \begin{bmatrix} 1.0 & 0.7 & 0.0 \\ 0.0 & 0.0 & 0.0 \\ 0.0 & 0.7 & 1.0 \end{bmatrix}.$$

Other improvements have been proposed as well [10,11]. (iii) The non-deterministic aspect of the Huang method could be solved by running the algorithm multiple times and choosing the image version with the highest log-likelihood [18]. However, the Huang method has the tendency to map confusion colors, whose differences are perfectly visible for normal-sighted people, to colors that become identical even for normal-sighted people (cf. Fig. 10f). Also, the interpolation of the daltonized image is not always as smooth as the authors claim (cf. the background color in Fig. 10b, d and e and the face color in Fig. 10f). (iv) Lastly, the differentiation of colors in the Kuhn images are significantly reduced for normal-sighted people since they lie within the reduced color gamut of the color-deficient (cf. Fig. 3f). Thus, the increased contrast of confusion colors compares to a decrease in general color contrast for the normal-sighted.

In future work, we will check if the results of our proposed method echo in a psychometric scaling experiment based on the law of comparative judgment proposed by Thurstone [42]. We will use a pairwise comparison experiments [43, Chapter 8] investigating naturalness, pleasantness, general preference, etc. Then, we will analyze the correlation between preference and behavioral performance. However, a psychometric scaling experiment would require significantly more comparisons since each daltonized version has to be compared to each of the remaining versions. Thus, a psychometric scaling experiment would be much more resource-intensive.

## 6. Conclusion

We proposed a behavioral method to evaluate daltonization methods based on the visual-search paradigm for natural images and Ishihara plates. We could show, on the one hand, that the accuracy data reveals statistically significant differences between individual observer groups and the investigated daltonization methods. On the other hand, the response time data does not help in the analysis of the individual daltonization methods. The Kotera and Fidaner methods perform best among the chosen methods since they manifest improvement in behavioral response for the color-deficient and no deterioration for the normal-sighted. More so, the Kotera method leads to the best, improvement for deutan color-deficient observers. The Huang and the Kuhn methods are ranked lowest among the chosen daltonization methods. They do not manifest improvement in behavioral response for the color-deficient, but deterioration for the normal-sighted. More precisely, the Kuhn method leads to the most deterioration for normal-sighted observers. With this paper, we present a proof-of-concept for our proposed method. In future work, we will evaluate more daltonization methods with more observers including protan color-deficient observers.

## Acknowledgments

We thank Dr. Peter Nussbaum, Dr. Reiner Eschbach and Dr. Jonny Nersveen (all Gjøvik University College) for constructive feedback on the article, and Prof. Bruno Laeng (University of Oslo) for his help during the behavioral experiment and the data analysis. Also, we thank Manuel Oliveira and Jia-Bin Huang for providing the implementation of the Kuhn and the Huang methods respectively. This research has been funded by the Research Council of Norway through Project No. 221073 "HyPerCept – Colour and quality in higher dimensions".

## References

[1] G. Wyszecki, W. Stiles, Color Science, 2nd ed., John Wiley & Sons, Inc., 2000.
[2] A. Valberg, Lys Syn Farge, 1st ed., Tapir Forlag, 1998.
[3] E. Hansen, Fargeblindhet, 1st ed., Gyldendal Norsk Forlag AS, 2010.
[4] C. Rigden, 'The Eye of the Beholder' – designing for colour-blind users, Brit. Telecommun. Eng. 17 (1999) 2–6.
[5] S. Ishihara, Tests for Colour-Blindness – 24 Plates, 24 Plates ed., Kanehara Shuppan Co., Ltd., 1972.
[6] B.L. Cole, K.-Y. Lian, C. Lakkis, The new Richmond HRR pseudoisochromatic test for colour vision is better than the Ishihara test, Clin. Exp. Optometry 89 (2) (2006) 73–80.
[7] D. Farnsworth, The Farnsworth Dichotomous Test for Color Blindness: Panel D-15, Psychological Corporation, 1947.
[8] P. Lanthony, The desaturated panel D-15, Doc. Ophthalmol. 46 (1) (1978) 185–189.
[9] W. Nagel, Zwei apparate für die augenärztliche funktionsprüfung, Z. Augenh. 17 (3) (1907) 201–222.
[10] C.-N. Anagnostopoulos, G. Tsekouras, I. Anagnostopoulos, C. Kalloniatis, Intelligent modification for the daltonization process of digitized paintings, in: The 5th Int. Conference on Computer Vision Systems, Universität Bielefeld, Bielefeld, 2007.
[11] H.-J. Kim, J.-Y. Jeong, Y.-J. Yoon, Y.-H. Kim, S.-J. Ko, Color modification for color-blind viewers using the dynamic color transformation, in: IEEE International Conference on Consumer Electronics (ICCE), IEEE, 2012, pp. 602–603.
[12] J.-B. Huang, Y.-C. Tseng, S.-I. Wu, S.-J. Wang, Information preserving color transformation for protanopia and deuteranopia, IEE Signal Process. Lett. 14 (10) (2007) 711–714.
[13] G.M. Machado, M.M. Oliveira, Real-time temporal-coherent color contrast enhancement for dichromats, Computer Graphics Forum, vol. 29, Wiley Online Library, 2010, pp. 933–942.
[14] F. Viénot, H. Brettel, J.D. Mollon, Digital video colourmaps for checking the legibility of displays by dichromats, Color Res. Appl. 24 (4) (1999) 243–252.
[15] H. Brettel, F. Viénot, J.D. Mollon, Computerized simulation of color appearance for dichromats, J. Opt. Soc. Am. A 14 (10) (1997) 2647–2655.
[16] H. Kotera, Optimal daltonization by spectral shift for dichromatic vision, in: 20th IS&T Color and Imaging Conference (CIC20), Society for Imaging Science and Technology (IS&T), Los Angeles, CA, USA, 2012, pp. 302–308.
[17] J.-B. Huang, C.-S. Chen, T.-C. Jen, S.-J. Wang, Image recolorization for the colorblind, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2009 (ICASSP 2009), IEEE, 2009, pp. 1161–1164.
[18] J.-B. Huang, Why the Huang Daltonization Method is Non-Deterministic, Private Email Communication, June 27th 2015.
[19] G.R. Kuhn, M.M. Oliveira, L.A. Fernandes, An efficient naturalness-preserving image-recoloring method for dichromats, IEEE Trans. Visual. Comput. Graph. 14 (6) (2008) 1747–1754.
[20] J.E. Mazur, Learning and Behavior, sixth ed., Pearson Education Inc., Upper Saddle River, NJ, USA, 2005.
[21] A.M. Treisman, G. Gelade, A feature-integration theory of attention, Cognitive Psych. 12 (1) (1980) 97–136.
[22] J. Simon-Liedtke, J.Y. Hardeberg, Task-based accessibility measurement of daltonization algorithms for information graphics, in: 12th Congress of the International Colour Association (AIC 2013), Newcastle, UK, 2013, pp. 108–111.
[23] The Center for Universal Design [online], About Universal Design, 2008. <http://www.ncsu.edu/ncsu/design/cud/about_ud/about_ud.htm> (last checked: December 2nd, 2014).
[24] The Center for Universal Design [online], The Principles of Universal Design, 1997. <http://www.ncsu.edu/ncsu/design/cud/about_ud/udprinciplestext.htm> (last checked: December 2nd, 2014).
[25] J.T. Simon-Liedtke, I. Farup, B. Laeng, Evaluating color deficiency simulation and daltonization methods through visual search and sample-to-match: SaMSEM and ViSDEM, in: Color Imaging XX: Displaying, Processing, Hardcopy, and Applications, Proceedings of IS&T/SPIE Electronic Imaging, vol. 93, International Society for Optics and Photonics, San Francisco, CA, USA, 2015. 939513-939513-9.
[26] M. Pedersen, N. Bonnier, J.Y. Hardeberg, F. Albregtsen, Attributes of image quality for color prints, J. Electron. Imag. 19 (1) (2010) 01101601–01101613.
[27] J.T. Simon-Liedtke, I. Farup, Using a behavioral match-to-sample method to evaluate color deficiency simulation methods, (submitted for publication).
[28] E.B. Wilson, Probable inference, the law of succession, and statistical inference, J. Am. Stat. Assoc. 22 (158) (1927) 209–212.
[29] G.G. Løvås, Statistikk for universiteter og høgskoler, second ed., Universitetsforlaget, Oslo, Norway, 2008.
[30] A.M. Mood, Introduction to the Theory of Statistics, McGraw-hill, 1950.
[31] R. McGill, J.W. Tukey, W.A. Larsen, Variations of box plots, Am. Stat. 32 (1) (1978) 12–16.
[32] J. Peirce, PsychoPy Documentation, 2014. <http://www.psychopy.org/index.html> (last checked: 04/23/2015).
[33] NumPy Developers, NumPy Documentation, 2013. <http://www.numpy.org/> (last checked: 04/23/2015).

[34] SciPy Developers, SciPy Documentation, 2013. <http://www.scipy.org/> (last checked: 04/23/2015).
[35] Matplotlib Development Team, Maplotlib Documentation, 2014. <http://matplotlib.org/> (last checke: 04/23/2015).
[36] PyData Development Team, Pandas Documentation, 2012. <http://pandas.pydata.org/> (last checked: 04/23/2015).
[37] J. Simon-Liedtke, I. Farup, Empirical disadvantages for color-deficient people, in: Mid-Term Meeting of the International Colour Association (AIC 2015), International Colour Association, Tokyo, Japan, 2015, pp. 391–394.
[38] I. Bramão, L. Faísca, C. Forkstam, F. Inácio, S. Araújo, K.M. Petersson, A. Reis, The interaction between surface color and color knowledge: Behavioral and electrophysiological evidence, Brain Cognition 78 (1) (2012) 28–37.
[39] I. Bramão, L. Faísca, K.M. Petersson, A. Reis, The contribution of color to object recognition, in: Advances in Object Recognition Systems, InTech, 2012, pp. 73–88.
[40] I. Bramão, A. Reis, K.M. Petersson, L. Faísca, The role of color information on object recognition: a review and meta-analysis, Acta Psychol. 138 (1) (2011) 244–253.
[41] J. Simon-Liedtke, Colorama: extra color sensation for the color-deficient with gene therapy and modal augmentation, in: Mid-Term Meeting of the International Colour Association (AIC 2015), International Colour Association, Tokyo, Japan, 2015, pp. 1329–1332.
[42] L.L. Thurstone, A law of comparative judgment, Psychol. Rev. 34 (4) (1927) 273–286.
[43] P.G. Engeldrum, Psychometric Scaling: A Toolkit for Imaging Systems Development, Imcotek Press, 2000.