Research Article Vol. 38, No. 9 / September 2021 / Journal of the Optical Society of America A 1349

Journal of the Optical Society of America A

OPTICS, IMAGE SCIENCE, AND VISION

On the evaluation of temporal and spatial stability of color constancy algorithms

MARCO BUZZELLI* AND ILARIA ERBA

Department of Informatics, Systems, and Communication, University of Milano–Bicocca, 20126 Milan, Italy
**Corresponding author: marco.buzzelli@unimib.it*

Received 22 June 2021; revised 2 August 2021; accepted 10 August 2021; posted 11 August 2021 (Doc. ID 434860); published 25 August 2021

Computational color constancy algorithms are commonly evaluated only through angular error analysis on annotated datasets of static images. The widespread use of videos in consumer devices motivated us to define a richer methodology for color constancy evaluation. To this extent, temporal and spatial stability are defined here to determine the degree of sensitivity of color constancy algorithms to variations in the scene that do not depend on the illuminant source, such as moving subjects or a moving camera. Our evaluation methodology is applied to compare several color constancy algorithms on stable sequences belonging to the Gray Ball and Burst Color Constancy video datasets. The stable sequences, identified using a general-purpose procedure, are made available for public download to encourage future research. Our investigation proves the importance of evaluating color constancy algorithms according to multiple metrics, instead of angular error alone. For example, the popular fully convolutional color constancy with confidence-weighted pooling algorithm is consistently the best performing solution for error evaluation, but it is often surpassed in terms of stability by the traditional gray edge algorithm, and by the more recent sensor-independent illumination estimation algorithm.

https://doi.org/10.1364/JOSAA.434860

1. INTRODUCTION

Color constancy is defined as the human ability to perceive the chromatic appearance of a scene as relatively constant, notwithstanding changes in illumination conditions [1]. Inspired by this feature of the human vision system, the field of computational color constancy was born and developed with the objective to transfer the same ability to digital camera sensors. Computational color constancy, from now on referred to as simply "color constancy," is often modeled with two explicit steps: illuminant estimation and illuminant correction. The correction step in particular is often carried out through a von Kries-like transform [2], using a diagonal matrix to apply independent correction to the response of cone photoreceptors. Although known to be suboptimal and unable to fully handle metameric effects [3], the von Kries transform is commonly adopted due to its simplicity. Color constancy methods are usually compared to angular error metrics such as the recovery error [4] and the reproduction error [5]. The comparison of color constancy methods based on angular errors is sometimes aided by statistical tools such as the Wilcoxon test [6], or by graphical tools such as the Angle-Retaining Chromaticity (ARC) diagram [7]. For many years, angular error evaluation has been extremely useful in assessing color constancy methods and guiding the research. However, it neglects other important aspects of the characterization of color constancy algorithms, related to their stability in the video domain.

This property has become extremely valuable since consumer devices are increasingly used for video acquisition and reproduction [8]: In this context, the discomfort of poor illuminant correction is potentially amplified if such correction also changes over time without justification, thus introducing unpleasant flickering artifacts. To this extent, existing works have tackled the problem of temporally aware color constancy [9,10], exploiting the information coming from multiple frames to produce a more robust illuminant estimation. Nonetheless, traditional single-frame methods can also be applied to video sequences, with or without the aid of temporal consistency post-processing [11]. As such, the main goal of this investigation is to study the direct applicability of single-frame color constancy algorithms in the video domain. We have identified two possible scenarios of interest: moving subjects in front of the camera and a panning/zooming camera, as depicted in Fig. 1. In these cases, if the scene illuminant remains constant, the expected behavior of a color constancy algorithm is that the output is also constant, ignoring the intrinsic chromaticity of newly framed elements.

The two scenarios of interest can be analyzed by resorting to appropriately annotated datasets for video color constancy, the best to date being the Gray Ball dataset [12], and the very recent Burst Color Constancy (BCC) dataset [10]. Both scenarios are depicted in such datasets, and often co-occur in the same video sequence. We will thus refer to the general term temporal stability as the capability of a given algorithm to maintain a consistent

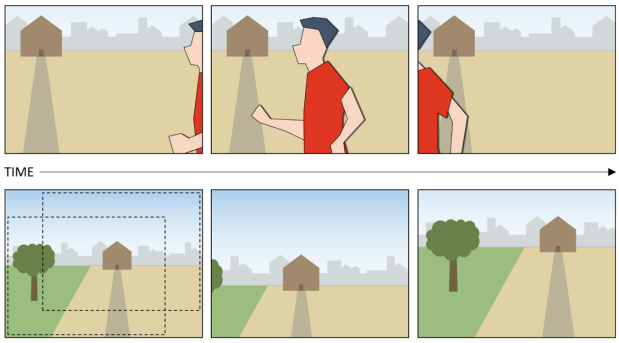

Fig. 1. Changing scene content under constant illuminant conditions: moving subjects in front of the camera (top) and panning/zooming camera (bottom). A color constancy algorithm that is temporally and spatially stable should provide a self-consistent response in these scenarios.

output response in a situation where the illuminant remains relatively constant, while the scene content changes over time. In addition, the specific panning/zooming camera scenario can be synthetically recreated by introducing cropping operations at various scales in individual frames of a color constancy dataset. This is similar to a preprocessing technique proposed by Qian *et al.* [9] to simulate a camera movement for robust BCC. This type of analysis on synthetic data allows us to provide more precise information about the specific case of a moving camera with stationary content. Given the nature of the experiment, we will refer to spatial stability as the capability of an algorithm to maintain a consistent illuminant estimation, assuming a unique illumination source, when looking at different portions of the scene.

For a real-life application, the assumption of unique illumination source is seldom completely verified because, for example, of the presence of multiple lights at different correlated color temperatures, mutual surface inter-reflections, or coexistence of sun/shadow areas. Nonetheless, we argue that a time-consistent correction of the image color can be a desirable property, as long as moderate camera movements are in action. In other words, spatial stability should not be expected, nor enforced, for drastic changes in image framing. For this reason, we apply our synthetic panning and zooming operations at various scales, we show the trending stability at all levels, and eventually focus our conclusions at the highest scale, which corresponds to the most moderate camera movement. Many algorithms for single-frame color constancy internally perform a spatially varying illuminant estimation, which is eventually mapped back to a global illuminant by consensus [13], or clustered to perform multi-illuminant estimation [14]. Considerable effort also has been made by the scientific community to provide spatially varying annotation of illuminant information, resorting to moving color targets such as with the DRONE dataset [15] or to computer-generated imagery such as with the MIST dataset [16].

In this paper we first analyze the aforementioned temporal and spatial stability properties of existing suitable datasets for computational color constancy. We then present a methodology to analyze the stability of color constancy algorithms themselves. Our goal is to define a common procedure for method comparison, and to assess the degree to which such methods are sensitive to changes in the scene that do not depend on the illuminant source. The information emerging from our analysis identifies methods that are intrinsically stable, thus holding the greatest potential for expansion to video color constancy without heavily relying on temporal consistency post-processing techniques.

2. DATASETS SELECTION AND PREPROCESSING

To perform the stability analysis of color constancy algorithms free of any bias from the underlying data, it is necessary to exploit datasets that are, respectively, temporally and spatially stable. In this section, we verify whether this condition is met on existing datasets and, when it is not, we describe the required preprocessing steps. The same procedure could be potentially applied to any future datasets for video color constancy that is properly annotated.

A. Gray Ball Dataset

Gray Ball [12] is one of the few datasets potentially suitable for temporal color constancy. It contains 11,346 images divided into 15 sequences, with many shots acquired at close intervals to one from another. Many of the images depict people, and include both indoor and outdoor scenarios, the latter taken in two different locations. The dataset was collected using a Sony VX-2000 digital video camera, and every shot includes the eponymous gray ball color target for ground truth annotation in the bottom-right corner. For illuminant estimation, the images have been masked to exclude the color target starting from pixel row 135 and column 226. The 360×240 px images are provided in nonlinear 8-bit RGB format. Since several color constancy methods rely on the assumption of linear sensors, the following pipeline has been applied:

1. Linearize the image (gamma correction with $\gamma = 2.2$).
2. Estimate the illuminant.
3. Delinearize the estimated illuminant ($\gamma = \frac{1}{2.2}$).

It should be noted that the precise value for gamma correction is derived by common usages of the Gray Ball dataset [17], but it is not guaranteed to match the actual device characterization. This linearization strategy, despite being an approximation for color constancy outside the camera pipeline [18], still allows the processing of images that are closer to the RAW sensor data with respect to the original sRGB, while at the same time performing error analysis between the output of unaltered existing methods and the official dataset ground truth.

The Gray Ball dataset does not respect the temporal stability characteristics needed for this work; therefore, it requires a specific preprocessing to remove temporally unstable sequences. In this paper, we refer to a temporally stable sequence as a sequence that (1) does not contain video cuts, (2) does not involve abrupt illuminant changes, and (3) does not span a wide set of illuminants (even if gradually changing). We address these three conditions in three different ways.

The video cuts have been resolved by human selection, meaning that the 15 original sequences of the Gray Ball dataset have

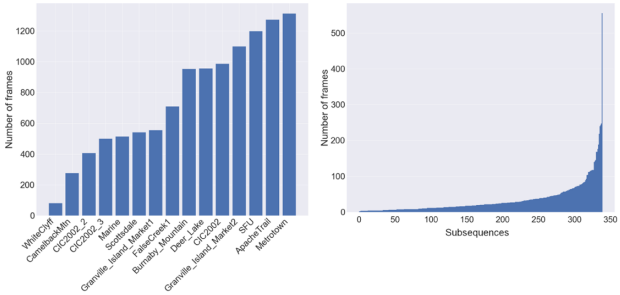

Fig. 2. Visualization of the number of frames for each sequence of the Gray Ball dataset before the manual division for video cuts (left) and after (right).

been manually divided into 337 smaller sequences, containing only smooth transitions of the scene content. The final distribution of the resulting sequence lengths is shown in Fig. 2.

The identification of abrupt illuminant changes has been achieved by quantifying the largest change in the expected illuminant $E = (r, g, b)$ between consecutive frames. More precisely, for each pair of consecutive frames in a sequence S, we calculated the recovery error between their ground truth illuminants and then we selected the maximum of such errors. From now on we will refer to this metric as the maximum illuminant change (MIC):

$$MIC(S) = \max(\mathrm{err}_{\mathrm{rec}}(E_{S_i}, E_{S_{i+1}})), \quad I = 1 \dots N_S - 1, \tag{1}$$

where N_S is the number of frames of sequence S.

The recovery error [4] as used in Eq. (1), is computed between two generic illuminants U and V as

$$\mathrm{err}_{\mathrm{rec}}(U, V) = \arccos\left(\frac{U \cdot V}{|U||V|}\right), \tag{2}$$

where "·" indicates the dot product, and "||" the Euclidean norm.

To identify sequences spanning a large range of illuminants, we instead relied on a metric for scatteredness. Specifically, we first converted the ground truth illuminants into ARC [7], a bidimensional representation where Euclidean distances correspond to angular distances in the original RGB space. Then, we computed the standard distance [19] of the resulting points, which is a bidimensional generalization of the standard deviation, defined as

$$STD(S) = \sqrt{\sum_{i=1}^{N_S} \frac{(x_{S_i} - \overline{x_S})^2}{N_S} + \sum_{i=1}^{N_S} \frac{(y_{S_i} - \overline{y_S})^2}{N_S}}, \tag{3}$$

where (x_{S_i}, y_{S_i}) are the ARC coordinates of the i-th illuminant of sequence S, and $(\overline{x_S}, \overline{y_S})$ indicates the average of each coordinate for the sequence.

The information captured by these measures is visualized in Fig. 3: For each sequence, we show the illuminant change between consecutive frames (whose maximum corresponds to MIC), as well as the ground truth illuminants distribution in ARC (whose scatteredness corresponds to STD). The two metrics, MIC and STD, were then combined to provide a single value that describes the instability of each sequence: We first computed the standard score of both metrics by normalizing them for the corresponding cross-sequence average and standard deviation, and we subsequently computed an equal-weight average. The resulting distribution was finally split in half using the median value as a threshold to divide the dataset into 168 stable sequences and 169 unstable sequences. In the following, we will refer to the selection of temporally stable sequences as the filtered Gray Ball dataset. Backtracking this division to the initial measures, it roughly corresponds to applying a threshold over MIC at 1.5° and STD at 0.8°, which appears adequate after a visual inspection of the dataset. However, due to the arbitrary nature of any specific threshold, we make available for public download our entire dataset division into subsequences, along with the corresponding values of stability-related measures to allow further developments by other researchers [20].

With respect to spatial stability, the Gray Ball dataset has been used for many years for global illuminant estimation analysis, under the implicit assumption of spatial stability. This assumption is, however, generally unsubstantiated. It is possible, for example, to find outdoor scenarios where part of the scene is

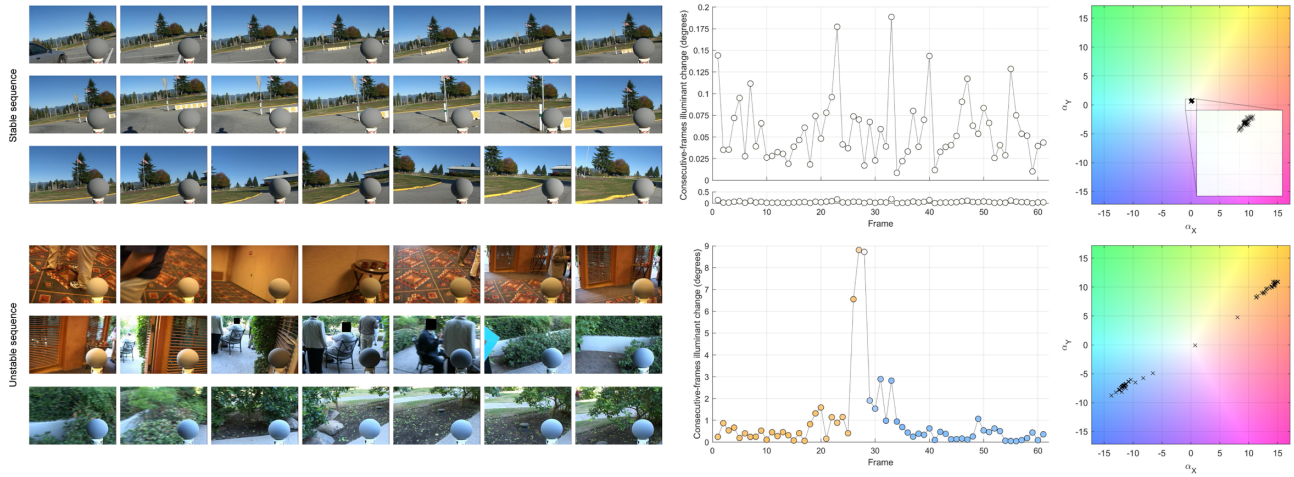

Fig. 3. Temporal stability analysis of two sequences from the Gray Ball dataset: a stable sequence ($MIC = 0.189$, $STD = 0.115$), top; and an unstable sequence ($MIC = 8.812$, $STD = 14.912$), bottom. For each sequence, we show a sample of the frames (left), the illuminant change between consecutive frames (center), and the illuminants distribution in ARC diagram (right).

illuminated by direct sunlight and part of it is in shadow, illuminated only by the blue of the sky. This configuration breaks the assumption of constant illumination across the image and, furthermore, it cannot be automatically filtered. Despite the fact that the color target (the gray ball) can capture incoming light from different directions, it only describes the illumination condition in the foreground of the picture, and it does not provide any way to associate the different illuminant chromaticities to specific image regions. Notwithstanding these considerations, for the sake of completeness and for consistency with the existing corpus of color research, we also will perform spatial stability analysis on this dataset as well.

B. BCC Dataset

The BCC dataset, sometimes referred to as the Temporal Benchmark dataset, was recently presented by Qian *et al.* [10], and it was specifically collected to meet the requirements of the temporal color constancy problem. It consists of 600 sequences of varying length (between 3 and 17 frames), divided in 400 sequences for the training set and 200 for test set, the latter used in our analysis. Consistent with the Gray Ball dataset, BCC covers indoor and outdoor scenes with varying weather and daylight conditions. The images were shot with the use of a Huawei Mate 20 Pro mobile phone, and stored in a proprietary 16-bit RAW format. Reprocessed 8-bit PNG images at 3648×2736 px resolution were also made available by the dataset authors, and these images were specifically used in our work.

With respect to the temporal color constancy problem, the sequences collected for the BCC dataset are assumed implicitly stable by design. This allowed the authors to avoid capturing the images with a color calibration target installed in the scene, and thus to avoid unintentionally conveying information to learning-based methods. Instead, the SpyderCube calibration target was put in the scene immediately after the sequence acquisition to create an out-of-sequence reference shot that represents the entire video sequence. For these reasons, there has been no need, nor possibility, of a preprocessing step from our part.

Concerning the spatial stability, the BCC dataset was collected and presented without any explicit statement in terms of single or multiple illuminant sources. A visual inspection of the dataset images confirmed the absence of images visibly illuminated by multiple sources of light, with the exception of few daylight/shadow instances, as also observed in the Gray Ball dataset.

3. ANALYZED COLOR CONSTANCY METHODS

We selected a variety of color constancy algorithms for our stability analysis, including traditional solutions based on handcrafted features, as well as more recent approaches based on deep learning. All methods are sensor-independent and, when necessary, trained on different datasets than the ones used for our analysis to ensure the absence of any bias and to provide fair results. This particular set of methods has been selected as a case study; however, the same procedure can be applied to any existing method for color constancy.

Edge-based color constancy (EB) [21] is a popular framework introduced in 2007 by van de Weijer *et al.* as a generalization of multiple algorithms based on low-level image statistics. The free parameters of these methods (Minkowski norm p and standard deviation σ) have been selected as reported in [22]:

- Gray World (GW): $p = 1, \sigma = 0$.
- White Point (WP): $p = \infty, \sigma = 0$.
- Shades of Gray (SoG): $p = 4, \sigma = 0$.
- General Gray World (GGW): $p = 9, \sigma = 9$.
- 1st order Gray Edge (GE1): $p = 1, \sigma = 6$.
- 2nd order Gray Edge (GE2): $p = 1, \sigma = 1$.

The standard deviation parameter σ describes the Gaussian filter applied by the underlying algorithms, and as such its impact on the final performance is tightly related to the size of the input image. We downscaled the images from the BCC dataset to have the maximum side be 360 pixels long, thus reaching the same dimensions as the images from the Gray Ball dataset. Preliminary experiments also showed that downscaling the BCC dataset resulted, on average, in better performance with regard to upscaling the Gray Ball dataset. This pre-processing has been applied only for edge-based color constancy, since other algorithms have different requirements or involve an internal rescaling of the input image.

More recent color constancy algorithms have also been considered. Cheng *et al.* [23] introduced a color constancy algorithm based on principal component analysis (PCA), observing that the mere analysis of color distribution provides as much information for illuminant estimation as a more complex spatial analysis. Their solution selects a predefined percentage of dark and bright pixels using a projection distance in the color distribution. In our experiments, the percentage parameter has been set to 3.5% following the best-performing configuration reported by the authors.

The grayness index (GI) [14] is a learning-free metric developed by Qian *et al.* to identify neutral surfaces (gray pixels) in an input image following the dichromatic reflection model [24]. This allows the estimation of single illuminant as well as multiple illuminant information. The default pretuned parameters from the official implementation have been used in our work.

Quasi-unsupervised color constancy (QU) [25] was developed by Bianco *et al.* to detect achromatic pixels in color images, after conversion to gray scale. Their solution is based on a convolutional neural network that can be trained without color constancy annotation, relying instead on the weak assumption that training images have been approximately balanced. The model used in this analysis was trained on images from the ILSVRC2012 dataset of the ImageNet initiative [26].

Fully convolutional color constancy with confidence-weighted pooling (FC^4) [27] by Hu *et al.* implements a neural network architecture that assigns confidence weights to various patches of an input image, based on the level of information and reliability that such patches are estimated to carry for the task of color constancy. The official implementation is supplied with pretrained models on each fold of the ColorChecker dataset [28]. In our analysis, we used the SqueezeNet-based model [29] pretrained on "fold 2 and 0."

In sensor-independent illumination estimation (SIIE) [30] authors Afifi *et al.* developed a learnable sensor-independent pseudo-RAW space to be used to "canonicalize" the RGB values of any given camera, under the explicit assumption of

input linear RAW–RGB images. Due to the nature of the Gray Ball dataset, where images are not in RAW format but already processed by an undisclosed camera pipeline, this method is expected to underperform, despite our synthetic linearization. For our analysis, we used the Matlab 2018b model pretrained on the NUS [23] and Cube+ [31] datasets.

4. ANALYSIS OF TEMPORAL STABILITY

In this section, we assess the temporal stability of color constancy algorithms, under the assumption of temporally stable sequences, such as those from the filtered Gray Ball dataset and the BCC dataset.

We applied two measures to describe the temporal stability of a color constancy algorithm: maximum illuminant change MIC and standard distance STD. These are the same criteria defined in Eqs. (1) and (3) of Section 2 to automatically identify stable sequences in the Gray Ball dataset; however, in this case the evaluation has been performed on estimated illuminants as opposed to ground truth illuminants. Each method was assigned two temporal stability scores, by averaging each of the two aforementioned measures across the sequences, for any given dataset.

Temporal stability alone is hardly effective to evaluate the quality of a color constancy algorithm. For example, a "do nothing" algorithm would score the best value for temporal stability, while not being able to produce an effective illuminant estimation. For this reason, all algorithms have also been evaluated in terms of traditional single-frame error measures, such as the recovery and reproduction error. The recovery error is computed according to Eq. (2), while reproduction error is computed as

$$\mathrm{err}_{\mathrm{rep}}(U,\ V) = \arccos\left(\frac{\frac{U}{V}}{|\frac{U}{V}|\sqrt{3}}\right), \tag{4}$$

where U is the ground truth illuminant and V the estimated illuminant. For the sake of consistency with the stability measures, in this analysis, we averaged the error values for each frame in a sequence, and subsequently averaged the cross-sequence results. Our stability/error evaluation is conceptually equivalent to assessing a solution in terms of precision and accuracy. The results related to the filtered Gray Ball dataset are presented in Table 1, in terms of maximum illuminant change (MIC), standard distance (STD), recovery error ($\mathrm{err}_{\mathrm{rec}}$), and reproduction error ($\mathrm{err}_{\mathrm{rep}}$). The two temporal stability metrics are also visualized in Fig. 4 in conjunction with the recovery error to better visualize the performance of the analyzed methods.

From the experiment on the Gray Ball dataset, we first observe that the methods perform very similarly for both the standard distance and maximum illuminant change metrics. Generally speaking then, stability and accuracy are also partially correlated, as highlighted in Fig. 4. This behavior is a consequence of our focus on temporally stable datasets: With such data, a method can be globally accurate only if it is also temporally stable. This is specifically manifest in the absence of points in the top-left corner of the plots. Despite this correlation, several rank inversions are present between stability and error measures. For example, while FC^4 is the most accurate method, it is surpassed in MIC-based stability by several methods. Of these, GE2 and

Table 1. Temporal Stability and Error Evaluation of Color Constancy Algorithms on the Filtered Gray Ball Dataset[a]

	Stability Measures		Error Measures	
Method	$MIC \downarrow$	$STD \downarrow$	$\mathrm{err}_{\mathrm{rec}} \downarrow$	$\mathrm{err}_{\mathrm{rep}} \downarrow$
GW [21]	3.38	2.76	6.79	7.06
WP [21]	2.81	1.55	5.76	6.01
SoG [21]	3.14	2.57	5.88	6.06
GGW [21]	3.82	3.01	6.31	6.52
GE1 [21]	2.78	2.35	5.66	5.89
GE2 [21]	**1.80**	**1.39**	5.41	5.74
PCA [23]	3.66	2.74	5.75	6.04
GI [14]	7.81	4.95	7.65	8.03
QU [25]	2.89	2.15	5.40	5.63
FC^4 [27]	3.35	1.97	**4.63**	**4.97**
SIIE [30]	2.32	2.30	6.31	6.65

[a]All values are expressed in degrees; the lower the better. In Tables 1–4, the best values are displayed in bold.

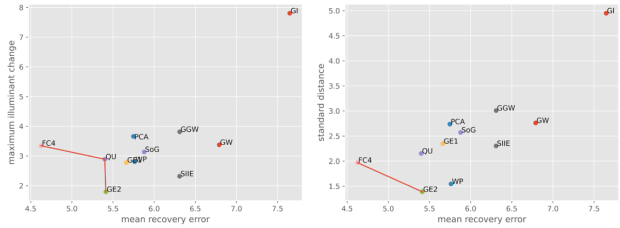

Fig. 4. Visualization of the temporal stability and recovery error of different methods on the Gray Ball dataset. Temporal stability is expressed as either maximum illuminant change (left) or as standard distance (right). The Pareto front is visualized. For all measures, the lower the better.

WP also appear to be the most stable in terms of the STD metric. Conversely, GI, the worst method on this dataset, performs consistently the worst on all the chosen metrics.

The same analysis for temporal stability has been performed on the BCC dataset, as illustrated in Table 2 and Fig. 5. Similar observations from the Gray Ball also can be extended to this dataset, in terms of the correlation between the different metrics. The SIIE method outperformed FC^4 on the Gray Ball dataset only in terms of MIC, while it performs consistently better for both temporal stability metrics on the BCC dataset. In this case, the single worst-performing method according to all metrics is the very simple white point (WP) algorithm.

The different conclusions that can be derived from analyzing the two datasets can be traced back to the type of images and the type of annotations. We recall that the Gray Ball is not distributed in linear RAW format, which limits the accuracy of the color constancy algorithms. We note the type of annotations because the BCC only has sequence-level ground truth information, which is in line with the assumption of temporal stability, but reduces the precision of the analysis.

5. ANALYSIS OF SPATIAL STABILITY

In this section, we assess the spatial stability of color constancy algorithms, under the implicit assumption of spatially stable datasets. We divided each image into five windows on a set of

Table 2. Temporal Stability and Error Evaluation of Color Constancy Algorithms on the BCC Dataset[a]

	Stability Measures		Error Measures	
Method	$MIC \downarrow$	$STD \downarrow$	$err_{rec} \downarrow$	$err_{rep} \downarrow$
GW [21]	2.70	3.20	5.32	7.08
WP [21]	6.31	5.07	6.98	8.26
SoG [21]	4.10	3.49	5.31	6.95
GGW [21]	4.37	4.01	5.92	7.61
GE1 [21]	3.58	3.40	5.52	7.30
GE2 [21]	3.40	3.01	4.79	6.10
PCA [23]	4.01	3.33	5.16	7.12
GI [14]	4.11	3.50	4.63	6.30
QU [25]	3.26	2.95	4.28	5.87
FC^4 [27]	3.08	2.64	**4.21**	**5.75**
SIIE [30]	**2.54**	**2.52**	4.49	6.06

[a]All values are expressed in degrees; the lower the better.

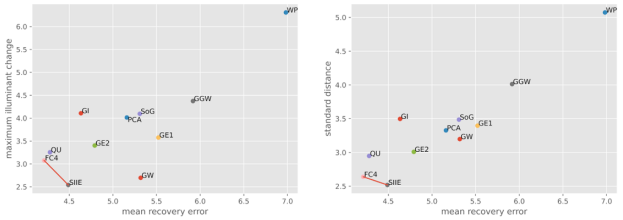

Fig. 5. Visualization of the temporal stability and recovery error of different methods on the BCC dataset. Temporal stability is expressed as either maximum illuminant change (left) or as standard distance (right). The Pareto front is visualized. For all measures, the lower the better.

fixed locations that cover the entire image: one for each angle and one in the center. A larger window size implies a higher overlap among windows, which corresponds to moderate camera movements in our synthetic setup. In such a scenario, we argue that having a consistent output in the color correction also is a desirable property for real-life applications because the change in overall incident illumination on such a scale can be expected to be limited. Given the arbitrary nature of fixing a window size, we present information at various scales from 50% to 90% of the original image sides, with a step of 10%. We will refer to the final scale (90%) to derive any conclusions about spatial stability.

All the analyzed color constancy methods have been applied to each window of each image. As for the temporal stability problem, we are interested in capturing both the accuracy and precision of analyzed algorithms; specifically, we resort to angular errors err_{rec} and err_{rep} for accuracy, and standard distance *STD* for precision. The maximum illuminant change *MIC* was well suited to highlight flickering phenomena in temporal sequences, but it does not provide any meaningful information if applied to the five windows of spatial stability analysis.

Table 3 and Fig. 6 report the aforementioned metrics on the Gray Ball dataset, with a detail per window size in the figure. Since larger windows are necessarily more overlapped, they are expected to lead to better stability, so any comparison should be made across methods and not across window sizes. This expectation is verified, as visible in the left side of Fig. 6, where all curves exhibit a monotonic decreasing behavior. The plotted information at a 90% window side is also presented numerically

Table 3. Spatial Stability Evaluation of Color Constancy Algorithms on the Gray Ball Dataset at 90% Window Side[a]

	Stability Measures	Error Measures	
Method	$STD \downarrow$	$err_{rec} \downarrow$	$err_{rep} \downarrow$
GW [21]	0.49	7.09	7.62
WP [21]	0.34	6.83	7.08
SoG [21]	0.44	6.17	6.49
GGW [21]	0.60	6.82	7.19
GE1 [21]	0.45	6.05	6.44
GE2 [21]	**0.28**	5.74	6.16
PCA [23]	0.49	6.40	6.85
GI [14]	0.96	7.02	7.61
QU [25]	0.55	6.45	6.73
FC^4 [27]	1.12	**5.67**	**6.02**
SIIE [30]	0.49	7.53	7.92

[a]All values are expressed in degrees, and the lower the better.

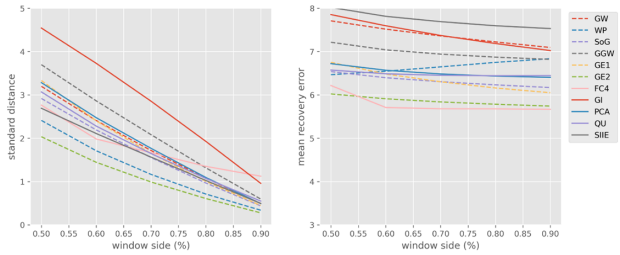

Fig. 6. Visualization of the standard distance (left) and recovery error (right) of different methods on the Gray Ball dataset. For all measures, the lower the better.

in Table 3. The most spatially stable algorithm on the Gray Ball dataset is GE2, in line with the analysis on temporal stability from Section 4. Interestingly, FC^4 is the least stable algorithm, specifically for large window sizes (90%), although it maintains a relatively consistent stability performance for smaller window sizes, compared to other methods. In terms of error analysis, a general trend of improvement with larger window sizes is also expected from the recovery error curves in Fig. 6. To this extent, the only exception appears to be WP, which exhibits a reverse trend: One possible explanation is that, by forcing the algorithm to ignore parts of the image, it is more likely to produce a better estimation from one or more windows. The behavior of FC^4 is also unusual, displaying a significant improvement from the 50% window size to the 60% size, but then essentially maintains the same error performance for the remaining window sizes. Nonetheless, it, in general, is consistently the best performing method with the Gray Ball dataset analysis.

Table 4 and Fig. 7 present the same spatial stability analysis on the BCC dataset, showing an overall comparable behavior. The expectation to observe improving performance with increasing window sizes also is respected for this dataset, with the only observed exception being the WP on the recovery error metric. FC^4 is the most accurate algorithm, with a significant gap from the second-best methods: QU and SIIE. In terms of spatial stability, however, the conceptually simple GW appears to exhibit the best performance.

Table 4. Spatial Stability Evaluation of Color Constancy Algorithms on the BCC Dataset at 90% Window Side[a]

Method	Stability Measures	Error Measures	
	$STD \downarrow$	$err_{rec} \downarrow$	$err_{rep} \downarrow$
GW [21]	**0.24**	5.65	7.46
WP [21]	0.55	6.70	7.99
SoG [21]	0.31	5.52	7.18
GGW [21]	0.40	6.07	7.77
GE1 [21]	0.38	5.88	7.72
GE2 [21]	0.29	5.03	6.38
PCA [23]	0.30	5.40	7.39
GI [14]	0.40	4.87	6.55
QU [25]	0.40	4.45	6.04
FC^4 [27]	0.43	**3.84**	**5.33**
SIIE [30]	0.29	4.55	6.10

[a]All values are expressed in degrees, and the lower the better.

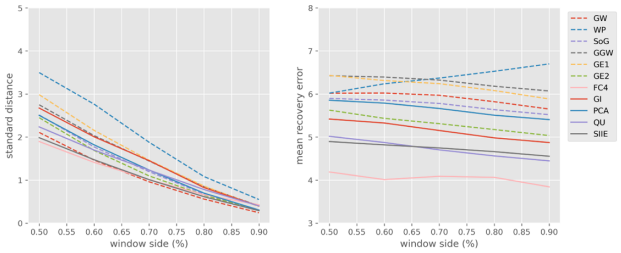

Fig. 7. Visualization of the standard distance (left) and recovery error (right) of different methods on the BCC dataset. For all measures, the lower the better.

It is also interesting to observe that, on the BCC dataset, the error statistics on learning-based algorithms can be neatly separated from those of traditional handcrafted solutions, although this does not apply to the Gray Ball dataset.

6. DISCUSSION OF AGGREGATED RESULTS

The analysis presented in this paper highlights the importance of evaluating algorithms according to multiple measures, and it can be useful in selecting the most appropriate color constancy method depending on specific application constraints.

To this extent, Table 5 presents an aggregated view of the information produced in the previous sections. The temporal stability rank is based on an average of temporal stability measures MIC and STD on both the Gray Ball and BCC datasets. Similarly, the spatial rank is based on the average of spatial STD on both datasets, and the error rank is based on err_{rec} and err_{rep} on the two datasets. Finally, a global rank is presented in the last column of Table 5. This ordering is determined through Borda's method for rank aggregation [32,33], where individual ranks are averaged, and elements are ranked again based on such average. This approach enables the combination of statistics coming from different domains (temporal, spatial, error), because it provides invariance to the different magnitude and distribution of the underlying values. An equal-weighted average has been used for our analysis; however, in the future different applications might motivate the selection of different

Table 5. Aggregated Ranks of the Analyzed Color Constancy Methods, According to Multiple Measures[a]

Method	Temporal Rank	Spatial Rank	Error Rank	Global Rank
GW [21]	5 (3.01)	2 (0.37)	9 (6.76)	6
WP [21]	10 (3.94)	7 (0.44)	11 (6.95)	9
SoG [21]	7 (3.33)	2 (0.37)	4 (6.19)	3
GGW [21]	9 (3.80)	9 (0.50)	10 (6.78)	9
GE1 [21]	6 (3.03)	6 (0.42)	7 (6.31)	7
GE2 [21]	1 (2.40)	1 (0.28)	3 (5.67)	1
PCA [23]	8 (3.44)	5 (0.40)	6 (6.26)	7
GI [14]	11 (5.09)	10 (0.68)	8 (6.58)	11
QU [25]	4 (2.81)	8 (0.47)	2 (5.61)	4
FC^4 [27]	3 (2.76)	11 (0.77)	1 (5.05)	5
SIIE [30]	2 (2.42)	4 (0.39)	5 (6.20)	2

[a]Underlying values are expressed in degrees.

weights for the temporal, spatial, and error components. In this particular setup, the best ranking method at a global level appears to be GE2, coherently with the individual temporal and spatial rank assessments. The second method is SIIE, which strikes a good balance across all evaluation criteria. The highly accurate and temporally stable method FC^4 is penalized in the global rank by its spatial instability, thus achieving intermediate overall performance. Finally, the lowest-ranked method in our experimental setup appears to be GI, which is negatively affected by its poor error performance on the Gray Ball dataset, and by its generally low stability.

7. CONCLUSIONS

We have introduced a new methodology to evaluate color constancy algorithms by taking into account their temporal and spatial stability. We have selected two color constancy datasets from the state of the art: the Gray Ball and the BCC, which we have analyzed and preprocessed for our evaluation, making the resulting characterization available for public download. We have conducted a case study on a wide set of color constancy algorithms, although our evaluation methodology can be applied to any given method. Concerning temporal stability, which measures the output consistency throughout frames in a video sequence, we have observed a general correlation with traditional error metrics, although some notable exceptions have been identified. The popular FC^4 algorithm, for example, is consistently the best performing one in terms of angular error, but it is outperformed in terms of stability by the SIIE algorithm on both analyzed datasets, and by several other methods on the Gray Ball dataset. The spatial stability analysis, which evaluates their output consistency across multiple windows of the input image, also led to similar conclusions: FC^4 has been identified as the least stable algorithm for large window sizes on the Gray Ball dataset, and is among the least stable ones on the BCC dataset, despite confirming its supremacy in terms of traditional error measures.

The analysis conducted in this paper provides the basis to identify those single-shot color constancy algorithms that have the greatest potential for expansion to video color constancy. Future investigations could also account for computational complexity: A method that is characterized by good or average

performance in terms of traditional angular error, but which displays scarce temporal and spatial stability, would potentially require a post-processing step to enforce temporal consistency. The resulting overhead at inference time could be prohibitive for a video-oriented application if the initial method is not inherently efficient.

Disclosures. The authors declare no conflicts of interest.

Data Availability. Data underlying the results presented in this paper are available in [20].

REFERENCES

1. D. H. Foster, "Does colour constancy exist?" Trends Cogn. Sci. **7**, 439–443 (2003).
2. J. von Kries, "Theoretische studien über die umstimmung des sehorgans," in *Festschrift der Albrecht-Ludwigs-Universität* (1902), pp. 145–158.
3. A. D. Logvinenko, B. Funt, H. Mirzaei, and R. Tokunaga, "Rethinking colour constancy," PLoS One **10**, e0135029 (2015).
4. S. D. Hordley and G. D. Finlayson, "Reevaluation of color constancy algorithm performance," J. Opt. Soc. Am. A **23**, 1008–1020 (2006).
5. G. D. Finlayson and R. Zakizadeh, "Reproduction angular error: An improved performance metric for illuminant estimation," Perception **310**, 1–26 (2014).
6. F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in Statistics* (Springer, 1992), pp. 196–202.
7. M. Buzzelli, S. Bianco, and R. Schettini, "ARC: Angle-retaining chromaticity diagram for color constancy error analysis," J. Opt. Soc. Am. A **37**, 1721–1730 (2020).
8. S. Wojcicki, "YouTube at 15: My personal journey and the road ahead," 2020. https://blog.youtube/news-and-events/youtube-at-15-my-personal-journey. Accessed on May 25, 2021.
9. Y. Qian, K. Chen, J. Nikkanen, J.-K. Kamarainen, and J. Matas, "Recurrent color constancy," in *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 5458–5466.
10. Y. Qian, J. Käpylä, J.-K. Kämäräinen, S. Koskinen, and J. Matas, "A benchmark for burst color constancy," in *European Conference on Computer Vision* (Springer, 2020), pp. 359–375.
11. W.-S. Lai, J.-B. Huang, O. Wang, E. Shechtman, E. Yumer, and M.-H. Yang, "Learning blind video temporal consistency," in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 170–185.
12. F. Ciurea and B. Funt, "A large image database for color constancy research," in *Color and Imaging Conference* (Society for Imaging Science and Technology, 2003), vol. **1**, pp. 160–164.
13. M. Buzzelli, R. Riva, S. Bianco, and R. Schettini, "Consensus-driven illuminant estimation with GANs," Proc. SPIE **11605**, 1160520 (2021).
14. Y. Qian, J.-K. Kamarainen, J. Nikkanen, and J. Matas, "On finding gray pixels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 8062–8070.
15. H. Aghaei and B. Funt, "A flying gray ball multi-illuminant image dataset for color research," in *Color and Imaging Conference* (Society for Imaging Science and Technology, 2020), vol. **2020**, pp. 142–149.
16. X. Hao and B. Funt, "A multi-illuminant synthetic image test set," Color Res. Appl. **45**, 1055–1066 (2020).
17. A. Gijsenij and T. Gevers, "Results per dataset (recovery error)," Color Constancy, 2011, http://colorconstancy.com/evaluation/results-per-dataset/index.html#sfugreyball_linear, accessed July 27, 2021.
18. M. Afifi, B. Price, S. Cohen, and M. S. Brown, "When color constancy goes wrong: Correcting improperly white-balanced images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 1535–1544.
19. J. E. Burt, G. M. Barber, and D. L. Rigby, *Elementary Statistics for Geographers* (Guilford, 2009).
20. M. Buzzelli and I. Erba, "Temporal and spatial stability of color constancy algorithms," Imaging and Vision Laboratory, 2021, http://www.ivl.disco.unimib.it/activities/color-stability/, accessed July 28, 2021.
21. J. Van De Weijer, T. Gevers, and A. Gijsenij, "Edge-based color constancy," IEEE Trans. Image Process. **16**, 2207–2214 (2007).
22. S. Bianco, C. Cusano, and R. Schettini, "Color constancy using CNNs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2015), pp. 81–89.
23. D. Cheng, D. K. Prasad, and M. S. Brown, "Illuminant estimation for color constancy: Why spatial-domain methods work and the role of the color distribution," J. Opt. Soc. Am. A **31**, 1049–1058 (2014).
24. S. A. Shafer, "Using color to separate reflection components," Color Res. Appl. **10**, 210–218 (1985).
25. S. Bianco and C. Cusano, "Quasi-unsupervised color constancy," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 12212–12221.
26. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," Int. J. Comput. Vis. **115**, 211–252 (2015).
27. Y. Hu, B. Wang, and S. Lin, "FC^4: Fully convolutional color constancy with confidence-weighted pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 4085–4094.
28. P. V. Gehler, C. Rother, A. Blake, T. Minka, and T. Sharp, "Bayesian color constancy revisited," in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2008), pp. 1–8.
29. F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5MB model size," arXiv:1602.07360 (2016).
30. M. Afifi and M. S. Brown, "Sensor-independent illumination estimation for DNN models," arXiv:1912.06888 (2019).
31. N. Banić, K. Koščević, and S. Lončarić, "Unsupervised learning for color constancy," arXiv:1712.00436 (2017).
32. J.-C. de Borda, "Mémoire sur les élections au scrutin," in *Histoire de l'Academie Royale des Sciences pour 1781* (1784).
33. S. Lin, "Rank aggregation methods," Wiley Interdiscip. Rev. Comput. Stat. **2**, 555–570 (2010).