Research Article | Vol. 27, No. 18 | 2 Sep 2019 | OPTICS EXPRESS 25611
Optics EXPRESS

Improving color constancy by selecting suitable set of training images

SHAO-BING GAO,[1] MING ZHANG,[2] AND YONG-JIE LI[2,*]

[1] College of Computer Science, Sichuan University, Chengdu 610065, China
[2] School of Life Science and Technology, Center for Information in BioMedicine, University of Electronic Science and Technology of China, Chengdu 610054, China
[*] liyj@uestc.edu.cn

Abstract: With very simple implementation, regression-based color constancy (CC) methods have recently obtained very competitive performance by applying a correction matrix to the results of some low level-based CC algorithms. However, most regression-based methods, e.g., Corrected Moment (CM), apply a same correction matrix to all the test images. Considering that the captured image color is usually determined by various factors (e.g., illuminant and surface reflectance), it is obviously not reasonable enough to apply a same correction to different test images without considering the intrinsic difference among images. In this work, we first mathematically analyze the key factors that may influence the performance of regression-based CC, and then we design principled rules to automatically select the suitable training images to learn an optimal correction matrix for each test image. With this strategy, the original regression-based CC (e.g., CM) is clearly improved to obtain more competitive performance on four widely used benchmark datasets. We also show that although this work focuses on improving the regression-based CM method, a noteworthy aspect of the proposed automatic training data selection strategy is its applicability to several representative regression-based approaches for the color constancy problem.

1. Introduction

The change of the light source color results in the change of the image color appearance captured by a camera [1, 2]. Such color shift may raise the difficulty to many computer vision tasks such as object recognition, tracking, and human pose estimation, all of which need to first discount the light source color and then retrieve the true color features of the images [3–6]. Many color constancy (CC) methods have been accordingly designed to solve this problem. According to whether or not involving the training procedure, the CC approaches can be divided into the statistic-based and learning-based methods.

For example, the gray-world-based algorithms [7–11] calculate the mean in each color channel of the image to estimate the illuminant. There are also other physically-based methods [12, 13], which introduce specific physical constraints (e.g., specular reflection) to estimate the scene illuminant. Recently, several CC algorithms that introduce various constraints of neural computation [14–18] have been proposed in order to build a framework of visual color constancy.

In contrast, learning-based methods often require extensive feature extraction and training [19–23], and in general perform quite better than statistic-based methods, but with the much higher computational cost. Recently, some relatively simple regression-based methods with the state-of-the-art performance have been proposed [24–27]. These learning-based methods work usually in the form of first extracting features and then finding a regression to capture the mapping between the features and the illuminant ground truth.

One of the representative regression-based CC methods is the so-called Corrected Moment (CM) illuminant estimation [28]. CM provides an efficient framework to integrate the illuminant estimation from multiple low level-based methods by using polynomial regression to find a

#369170 https://doi.org/10.1364/OE.27.025611
Journal © 2019 Received 7 Jun 2019; revised 13 Aug 2019; accepted 13 Aug 2019; published 26 Aug 2019

mapping between the features and the illuminant. Recently, based on the same rule, other more complicated regression techniques (e.g., regression tree [21] and deep learning [19, 29]) have been proposed to improve the learning based CC performance. We will call these algorithms as the CM-like methods in the following parts.

Although CM-like regression methods can currently produce the state-of-the-art performance among learning-based CC methods, the disadvantage of these methods is that they treat illuminant estimation as a black box without further considering the rule of image formation (e.g., similar scenes may have similar illuminants [30]). Moreover, CM applies a same correction matrix to all the test images without discriminating the intrinsic difference of various images. Thus, CM-like methods can not find an optimal model for each image according to its intrinsic property. We will fully discuss this in Section II. Moreover, we mainly focus on estimating the illuminant of the scene assuming that the illuminant is uniform over the entire scene in this work. For the scene with the varying illuminant, please refer to the excellent work on color constancy algorithms that try to estimate the illuminant on a per pixel basis [31–34].

The contributions of this work are as follows. According to the image formation model, we mathematically analyze one of the key factors (e.g., the novel feature $F^{'}$ introduced in this work) that influences the performance of regression-based CC. (2) A tenfold cross-validation based procedure is designed to automatically determine a proper size of the training subset for the test images. (3) A selection mechanism based on the feature $F^{'}$ is designed to adaptively pick a suitable subset of training data for each test image. (4) With this adaptive procedure, we build a general framework that can significantly improve the performance of many regression-based learning methods.

The related work is discussed in Section 2. The proposed method is presented in Section 3, in which we utilize the state-of-the-art CM to validate our proposed method. We then show the experimental results in Section 4. Finally, we conclude the work in Section 5.

2. Related work

There are mainly two ways to improve the performance of learning-based CC. The first way is to introduce more effective regression framework (e.g., deep learning), but with more complicated implementation. For example, Bianco et al. [19] used a Convolutional Neural Network (CNN) to build the mapping between the images and the illuminants. There are also other CNN-based CC methods adopting various strategies [20, 29, 35, 36]. Recently, to make deep-learning-based CC more understandable, Hu et al. [36] proposed a mechanism to reveal the confidence of deep CNNs in each region of the input images.

The regression-based CCs have typically the simple implementation and competitive performance [21, 28] compared to the classification-based CCs [37, 38]. The regression-based CCs evaluate their performance with the three-fold cross-validation procedure, i.e., randomly dividing the dataset into three parts. Then one part is selected as the test set and another two parts are left as the training set. Such procedure is repeated three times to make sure each part to be tested once. In one run of this procedure, all the images in the test set always share the same training set. However, considering each test image has its own intrinsic property, correcting each test image using a fixed correction matrix learned from a fixed training set may not work very well.

The second way to improve the performance of learning-based CC is to introduce the scene classification technique to select a suitable CC (e.g., grey edge [10]) to predict the illuminant according to the classification of each image [30, 37–39]. How to pick a suitable CC model according to the property of each test image has also been considered as a scene classification problem. These methods first group the training image set into several parts, and then train a particular CC model for each part. When a new test image comes, the CC model trained on the images that have the similar properties (e.g., similar SIFT features) to this test image will be used. For example, Gisenij et al. [37, 39] utilized the Weibull parameterization and MOG (Mixture of

Optics EXPRESS

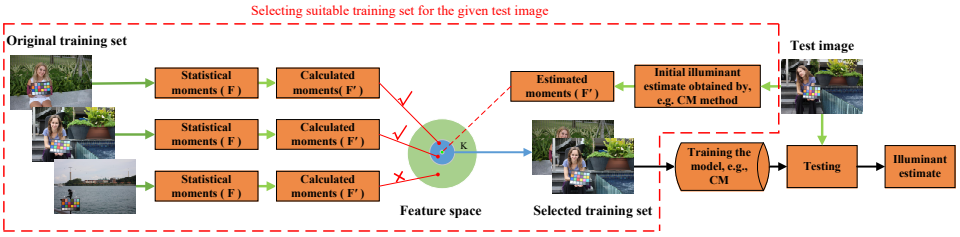

Fig. 1. The flowchart of the proposed method. For a given test image, some suitable images are selected from the original training set based on the proposed $F^{'}$ feature, which are used to train a specific regression based color constancy model (e.g., CM [28].)

Gaussian)-classifier to select a suitable CC method for each image. Bianco et al. [38] divided the images into indoor and outdoor scenes and selected the suitable CC methods accordingly.

Since these methods use the statistic-based CCs (e.g., grey edge [10]) to estimate the illuminant, their performance reaches no higher than the best of these low level-based algorithms for specific images. Moreover, these methods mentioned above classify the images based on the color distribution similarity of the retrieving images [30, 37, 40, 41] by assuming that the images with similar illumination have similar color distribution. Since similar or neighboring images are not always precisely affected by the similar illumination, thus the methods based on such assumption are not reasonable enough. Therefore, in order to further improve the CC performance, instead of selecting the statistic-based algorithms to estimate the illuminant, Wu et al. applied different learning-based CC algorithms for each class [42]. However, their method does not learn an optimal model for each image, but applies the same model to all the test images of each class. Exemplar-Based method [30] finds the nearest surface for the target surface, which also implicitly assumes that the similar images share the similar illuminants.

In short, all of the methods do not consider the fact that each image has its own particular properties [2], and thus should learn a correction model for each image. Moreover, which kind of features are more suitable to select the appropriate images to improve the training process has not been clearly explained in previous work. In this work, we introduce a selection mechanism to pick a suitable subset of training data for each test image using a novel illuminant related feature $F^{'}$, and the size of this training subset is automatically determined based on the whole training set. With these principled rules, a dynamic correction matrix is learned for each test image.

3. Proposed method

3.1. Image formation model

The response of a sensor to light and surface can be calculated as [43],

$$\rho^{E,S}(x) = \int_{\omega} R(\lambda)E(\lambda)S(\lambda, x)d\lambda \tag{1}$$

where $R(\lambda)$ denotes the camera sensor spectral sensitivity, $E(\lambda)$ and $S(\lambda, x)$ denotes respectively the spectral power distribution of the scene illuminant and the surface reflectance. The integral is taken over the visible spectrum ω. For practical application, if we define

$$\begin{aligned} \rho^{E} &= \int_{\omega} R(\lambda)E(\lambda)d\lambda \\ \rho^{S}(x) &= \int_{\omega} R(\lambda)S(\lambda, x)d\lambda. \end{aligned} \tag{2}$$

Equation (1) can be further simplified to a vector form [44],

$$\rho^{SE}(x) = \rho^{E} \cdot \rho^{S}(x) \tag{3}$$

where ρ^E denotes the scene illuminant vector that is commonly assumed to be uniform across the scene, $\rho^S(x)$ denotes the true surface color at location x viewed under a uniform white light source. Note that Eq. (3) generally holds in some sensor basis for a linear combination of the sensors [45].

We further abstract the regression-based method as follows,

$$\rho^E = F \cdot C \tag{4}$$

where F denotes the feature vector extracted from the original color-biased image and C denotes the regression mapping. Determined by the specific technique used, C can be a mapping matrix, a regression tree or a neural network.

3.2. Definition of feature moments

For a given N-pixel color-biased image, if we extract its feature vector F as a statistical moment of the image's RGB values, we have

$$F = moment(\rho^{SE}(x_1), \rho^{SE}(x_2), \cdots, \rho^{SE}(x_N)). \tag{5}$$

The simplest example of the $moment(.)$ is the global mean of per-channel (e.g., $\overline{R}, \overline{G}, \overline{B}$). For the per-channel moments, it is natural to further introduce the "cross moments" (e.g., $\overline{RG}^{0.5}$, indicating the square root of the global mean of the multiplication of R and G channels). We can further introduce the higher order moments. For a monomial of degree M and 3 variables the number of moments is equal to $\frac{(M+2)!}{2M!}$ [28].

In this work, we define the $moment(\cdot)$ as that of used in [28] for an image, i.e., the moments of a RGB image containing the first- and second-order items are calculated as

$$\begin{aligned} moment(\cdot) = (&\overline{R}, \overline{G}, \overline{B}, \overline{R^2}^{0.5}, \overline{G^2}^{0.5}, \overline{B^2}^{0.5}, \\ &\overline{RG}^{0.5}, \overline{RB}^{0.5}, \overline{GB}^{0.5}) \end{aligned} \tag{6}$$

where $\overline{R}$ represents the global mean of the R channel of the image, R^2 denotes to calculate the intensity's square of each pixel in the R channel, and $\overline{R^2}^{0.5}$ represents the square root of the global mean of R^2. Similar representations hold true for other channels. Hence, each element in $moment(\cdot)$ defined by Eq. (6) is a number and this moment is a feature vector with 9 dimensions, which is used to represent the features extracted from an image. Since the $moment(.)$ defined by Eq. (6) maintains the intensity-scaling property, which could well preserve the object colors and thus has been used for color correction and illuminant estimation [28,46].

Based on Eqs. (3) and (5), we can rewrite Eq. (4) as

$$\begin{aligned} \rho^E = moment(&\rho^E \cdot \rho^S(x_1), \rho^E \cdot \rho^S(x_2), \\ &\cdots, \rho^E \cdot \rho^S(x_N)) \cdot C. \end{aligned} \tag{7}$$

Due to the intensity-scaling property of the moment feature mentioned above, Eq. (7) can be further written as

$$\rho^E = \rho^E \cdot moment(\rho^S(x_1), \rho^S(x_2), \cdots, \rho^S(x_N)) \cdot C \tag{8}$$

if we further denote the moment term in Eq. (8) as

$$F' = moment(\rho^S(x_1), \rho^S(x_2), \cdots, \rho^S(x_N)) \tag{9}$$

then Eq. (8) could be further written as

$$\rho^E = \rho^E \cdot F' \cdot C. \tag{10}$$

Optics EXPRESS

Obviously, Eq. (10) teaches that for a regression-based illuminant estimation algorithm, the aim of training phase is to learn a mapping C so that the term $F^{'} \cdot C$ in the right side equals to an identity matrix. Thus, the performance of regression-based illuminant estimation is totally determined by how well the C can regress the moment $F^{'}$ of an image so that their multiplication equals to an identity matrix.

In other words, in order to find the optimal training images to learn the regression matrix C of a given test image, the training images should have exactly the same moment $F^{'}$ as the test image. Thus, the similarity of the moment $F^{'}$ between the test and the training images totally determines the final regression performance of illuminant estimation.

3.3. Calculation of feature moments

By comparing the expressions of Eqs. (4) and (10), we can get

$$F = \rho^{E} \cdot F^{'}. \tag{11}$$

Then, we can simply calculate the moment $F^{'}$ as

$$F^{'} = \rho^{E\dagger} \cdot F \tag{12}$$

where $\rho^{E\dagger}$ denotes the Moore-Penrose pseudo inverse of the scene illuminant vector ρ^{E}. The moment $F^{'}$ will be the key vector influencing the performance of a regression-based CC.

Equation (12) also teaches a fact that the final estimation of regression-based CC is affected not only by the surface reflectance (e.g., the statistical moment F), but also by the scene illuminant (e.g., the Moore-Penrose pseudo inverse of the scene illuminant vector $\rho^{E\dagger}$). The images having same surface reflectances may have various illuminants and their moment $F^{'}$ would be also different. Thus, it is not reasonable for these regression-based methods mentioned above [21,28,42,47] to use the same correction matrix to predict the illuminant for different test images. Moreover, previous CC algorithms [30,37,39] assume that the images with similar scenes hold similar illuminants and they estimate the illuminant through measuring the image similarity by using the general features (e.g., SIFT). However, our analysis indicates that we should use the illuminant estimation related moment $F^{'}$ to measure the similarity of two images, since the moment $F^{'}$ totally determines the performance of regression-based illuminant estimation.

Another contribution of our work is to introduce the statistical moment $F^{'}$, which can be used to help the selection of images with similar illuminants. It should be pointed out that, roughly, our proposed method is limited to the situation that the training set includes the illuminants that are similar to that of the test image. For example, it is difficult for our model to train using the images captured in the morning and estimate the illuminants for the test images taken in the nightfall, since these two conditions have very different illuminants.

Finally, most existing learning-based methods take the strategy that all the test images share the same training image set, which seems redundant since our model only needs some similar images (e.g., with similar illuminants) for each test image. Ideally, according to Eq. (12), if we can accurately find the moment $F^{'}$ of a test image from the training set, we can get the perfect illuminant estimate for this test image. However, in real situation the training images and the test images are absolutely separated and thus there are no exactly the same images in both of the training and test sets. Another problem is that in order to calculate the moment $F^{'}$, we need to know the illuminant ground truth for the test image. Obviously, this is not possible for our current problem since our task is to infer the illuminant for a test image without any information about the illuminant.

Thus, in order to avoid such a dilemma, we chose an alternative way to compute the moment $F^{'}$ for the test image. Since the existing learning-based methods can get relatively accurate illuminant estimation, we first utilize them to get a relatively accurate estimation of the moment $F^{'}$. Then, based on the estimation, the cosine distance of the moment $F^{'}$ vector between one

training image and one test image can be used to represent the similarity of the two images. The closer the distance between the two moments is, the more similar the illuminant and reflectance are between the training and the test images. Thus, for each test image, we can automatically pick the K most similar images from the original training dataset, or in other words, select the training images with the K top shortest cosine distances to the test image as the training subset for this test image, then learn a new illuminant prediction model based on the constrained subset of the training images. Figure 1 shows the framework of our proposed method.

3.4. Determination of the parameter K value

Our method introduces an very important parameter K, the number of the selected training images for a test image, and theoretically, the best K for each test image is different. However, in the following experiments, we will show that using a same parameter K for all the test images in a dataset can also greatly improve the performance. We firstly define the angular error (AE) as the metric to evaluate the error during the optimization of parameter K. The angular error ε between the estimated illuminant $\overline{e}$ and the illuminant ground truth e is computed as [48],

$$\varepsilon = cos^{-1}\left((\overline{e}\cdot e)/(\|\overline{e}\|\cdot\|e\|)\right). \tag{13}$$

Then, the proper K value in our paper is automatically determined according to the tenfold cross-validation based procedures listed below.

(1) The whole training dataset is randomly divided into ten parts, each of which contains N images.

(2) At each time, only one part is selected as the validating set, and the remaining nine parts as the temporary training set.

(3) Our method is ran for an exhaustive searching of parameter K on the temporary training set and the selected K is evaluated on the validating set based on the metric of AE, by which we can obtain a AE vs. K curve.

(4) The steps (2) and (3) are repeated 10 times with different validating sets at each time, by which we get 10 AE vs. K curves.

(5) Finally, we average the ten curves into one and choose the value of K with the lowest AE as the finally determined parameter K value.

To summarize the above procedures, we frame the criterion to estimate K as a cost function as follows. If $V(k)$ represents the AE of each validating image when selecting the k training images, our aim is to minimize the following cost function,

$$K = \arg\min_{k} \sum_{i=1}^{10}\sum_{j=1}^{N} V_{i,j}(k) \quad s.t. \quad 1 \le k \le 9\cdot N \tag{14}$$

where N is the image number in each part, "10" is the repeated times and "9" is the number of the training parts when executing the tenfold cross validation. It is clear that the parameter K is totally automatically determined only using the training set of a database, and this pre-determined K value will be applied on all the test images of this database.

3.5. About the CM and the generalization of our model

In this paper, we use CM [28,49] as the representative to verify the proposed strategy mentioned above. In the following experiments, we will show that the proposed method can significantly improve the performance of the original CM. It should be mentioned that besides CM, the CC method used to provide a initial illuminant estimation can be any CC method. The reason why we choose CM as the representative is primarily based on two considerations. On the one hand, CM holds nearly the state-of-the-art performance similar to that of all regression-based methods [19,21,28,42], but has very simple implementation. On the other hand, CM involves

Optics EXPRESS

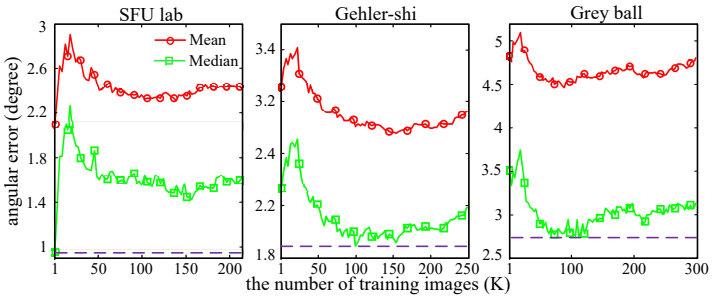

Fig. 2. The influence of the parameter K (i.e., the number of the selected similar images as training set) on the mean and median angular errors tested on the three different datasets.

the nonlinear cross terms in its extracted features, thus we can test our approach under a more complicated condition. In CM, the mapping between the image features and the illuminant is represented as a correction matrix that is used for all test images. Therefore, in our proposed strategy, the improved CM tries to find an optimal correction matrix for each image. In this work, in order to fairly compare the experimental results with CM, we list the results of the original CM reported in the literature [21, 28].

4. Experimental results

We evaluated the performance of our proposed strategy (selecting illuminant related moment features, SIRMF) based on selecting suitable images with the novel F' feature on four widely used benchmark datasets: the SFU lab dataset [25], the reprocessed version of the Gehler-Shi dataset [26, 50, 51], the Grey ball dataset [27], and the recent NUS dataset [24]. For all the comparisons, we only reported the best results of CM using the various number of edge moments for different datasets [21, 28] (e.g., 19-edge moment for the first three datasets, and 9-edge moment for the last dataset). Except that the AE defined in Eq. (13) is used as the metric to evaluate the performance of a CC algorithm, we also used the recently proposed reproduction angular error (RAE) to measure the illuminant estimate performance for each CC algorithm [44].

Similar to the recent literature [4], we used 3-fold cross validation to evaluate the proposed method on the SFU lab dataset, Gehler-Shi dataset, and NUS dataset. For Grey-ball dataset, since there is very strong correlation among images in this dataset, we used the training and test set splitted by Bianco et al. [38, 52] to avoid overfitting.

Our proposed SIRMF method contains the only parameter K (i.e., the number of the selected training images for a test image) that needs to be determined. Figure 2 shows the mean and median angular errors changing with the parameter K on the SFU lab dataset, the Gehler-Shi dataset, and the Grey ball dataset. For the SFU lab dataset, our method obtains the best illuminant estimation performance by just selecting the first most similar training image, since the illuminants of SFU lab dataset are primarily clustered within a small range. However, for the two natural image datasets (i.e., Gehler-Shi dataset and Grey ball dataset), we can observe that with the increasing of K, the angular error first increases and then decreases to the lowest point, and then increases again. The reason why we get the initial increase in Fig. 2 is due to the fact that we calculate the initial F' of test image based on the original estimation of CM, which is commonly not equal to the ground truth illuminant. Thus, the first increase is because of the insufficient filtering of training images, which means each image in the selected sub-training set can greatly affect the final estimation especially for those images with illuminant conditions that are quite different from the test image. With more images joining into the training set, the influence of those unrelated images could be relatively weakened, which finally leads to the lowest angular error. With the further increasing of K, more irrelevant training images would be again selected into the training

Optics EXPRESS

subset, which results in the second increasing of angular error. For each dataset, the reported results in the table are obtained using the same parameters (K) set at the point of lowest angular error (the dashed purple line in Fig. 2). Note that fixing the parameters on the whole dataset to report the performance of a CC method is a widely used strategy in CC community [4]. We report the results of method of the fixed parameter K in each table as SIRMF(K). For comparison, we also report the results with the automatically determined K in each table as SIRMF(auto).

Note that automatic selection of K happens only on the training set during the three-fold cross-validation. It is clear that there should be a K determined by the training set on each fold, and hence, three different K values should be reported, as shown in Table 1 (for the SFU lab dataset) and Table 3 (for the Gehler-Shi dataset). In details, for the SFU lab dataset, the three K values are all equal to 1, and for the Gehler-Shi dataset, the three K values are 86, 107 and 84, respectively. But for the grey ball dataset, since the training and test sets were originally splitted by Bianco et al. [38, 39], so we did not conduct the three-fold cross-validation, and there is only one K value reported in Table 4 for our SIRMF(auto).

In the experiments, we found that on each dataset the automatically determined optimal value of K using the principled way described earlier is quite close to the best one as shown in Fig. 2 (the dashed purple line in Fig. 2). For example, in Table 3 and Table 4 the automatically determined optimal K values on the Gehler-Shi dataset and Grey Ball dataset are respectively SIRMF($auto$, K=86, 107, 84) and SIRMF($auto$, K=110), which are quite close to the optimal K values on the whole dataset (SIRMF(K=100) and SIRMF(K=84)).

Basically, the illuminant estimation performances on the four datasets are significantly improved by selecting similar images from the training set for each test image to correspondingly learn a correction matrix using SIRMF. Since we aim to improve the performance of regression-based methods (e.g., CM), we mainly compared our proposed SIRMF with CM. Besides, we also compared with other recent state-of-the-art CC models [53].

4.1. SFU lab dataset

From Table 1, we can see that our model performs best in comparison to other CC algorithms on the SFU lab dataset. Particularly, our approach obtains an improvement of up to 52% over the original CM [28] in terms of AE's median. In order to understand that our method can find the most suitable training images to learn an optimal correction matrix for each test image, Fig. 3 illustrates five test images in this dataset under the condition of K=1, where K indicates the number of the selected most suitable images according to the cosine distance of the moment F'

Table 1. **AE and RAE of Various Methods on the SFU Lab Dataset.**

Methods	AE		RAE	
	Median	Mean	Median	Mean
WP [8]	6.48°	9.09°	7.4°	9.7°
GW [7]	7.00°	9.78°	7.5°	10.1°
GE2 [10]	2.74°	5.19°	3.0°	5.8°
PBG [54]	2.27°	3.70°	2.8°	4.2°
EBG [54]	2.3	3.9	2.7°	4.5°
SS [55]	3.45°	5.63°	–	–
WGE [11]	2.4	5.6	3.6°	6.1°
CM(9 edge) [28]	2.0°	2.6°	–	–
SIRMF(auto & K=1,1,1)	**0.96°**	**2.10°**	**1.09°**	**2.42°**

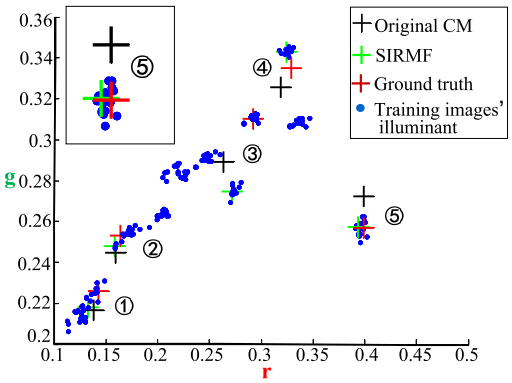

Fig. 3. The scatter plot of the illuminants of training images, illuminant ground truth of five test images, the corresponding illuminant estimates by CM, and the improved illuminant estimates by the proposed method.

between each training image and a given test image. Here, K=1 means that our proposed model only selects one most suitable training image to learn a new correction matrix for each test image on this dataset. Note that on this dataset, the optimal parameter of K determined by SIRMF(auto) is same to the best K value as densely searched in Fig. 2.

We can observe from Fig. 3 that for the group #2, #4 and #5, our method obtains the illuminant estimates that are much closer to the ground truth. Figure 4 visualizes the test images and the selected training images of the group #2, #4 and #3 by our proposed F' and SIFT. We can see that our proposed SIRMF model can select the training images with quite similar illuminants to the test images to learn a new correction matrix and thus can improve the accuracy of illuminant estimation. While SIFT can only select the images with similar scene contents but with quite different illuminants. It should be noted that in the first and second row of Fig. 4, although the test image (left) and the selected training image (middle) are different (please notice the little scene difference between the left image and the middle image in the first row), they indeed have the similar illuminants. Table 2 shows the corresponding AE trained using the selected images shown in middle and right columns of Fig. 4. It is clear that the performance of our proposed F' outperforms greatly than that of the SIFT in terms of the AE. The high AE of the SIFT in Table 2 teaches that selecting similar scene contents will lead to a great mistake under some circumstances.

SFU lab dataset includes 31 different objects under 11 different illuminant conditions, most images in this dataset with similar scene contents have quite different light source colors. Thus, particularly in this image dataset, selecting images with similar scene contents will lead to quite wrong illuminant estimate (e.g, the images in the second row of Fig. 4). Our proposed strategy is try to select the images with similar illuminants to help the re-training of correction matrix, instead of selecting the images with similar scene contents. Thus, evaluating our proposed model on SFU lab dataset is very challenging.

However, there is also the situation of group #3, for which the accuracy of illuminant estimation

Table 2. **AE of the Test Image on the SFU Lab Dataset in Fig. 4 by Using the Proposed Feature F' and the SIFT to Select the Training Images.**

Image	Group #2	Group #4	Group #3
F'	0.95°	0.99°	6.71°
SIFT	7.80°	10.95°	15.55°

Optics EXPRESS

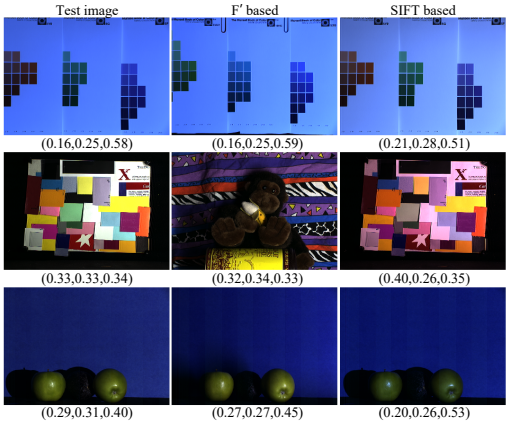

Fig. 4. The test and the selected training images of group #2, #4, and #3 in Fig. 3 by the features of F' and SIFT (from top to bottom : the images of group #2, #4, and #3). The numbers under each image denote the RGB components of the true light source color.

is reduced by the proposed F' based method. The reason is that we first use the illuminant estimation of the original CM to calculate the moment F', and we might get quite wrong estimation of moment F' if the illuminant estimated by CM has quite large error. Thus, we could not find the most suitable training images with similar illuminants to further learn an accurate correction matrix. For example, in the last row (group #3) of Fig. 4, although our method can select out the training image of similar scenes to the test image, they actually have different illuminants.

Therefore, to keep a stable improvement by our method, it is necessary to make a constraint stating that the new illuminant estimation can not deviate from the original illuminant estimation more than certain degree. In all the experiments on the four datasets tested here, we found that an angular error of 5° is an appropriate degree to measure the deviation between the new illuminant estimation by our strategy and the original illuminant estimation by CM. If the deviation is beyond this limitation, we will give up the new illuminant estimation and still use the original estimate obtained by CM.

4.2. Gehler-Shi dataset

The Gehler-Shi dataset contains 568 images captured by two digital cameras (CANON 1D and CANON 5D). For unbiased evaluation, the color checker utilized to calculate the ground truth in each image was masked out during the experiment as did by others [21, 28]. In Table 3, we show the results of several state-of-the-art methods on this dataset. The mean and median AEs of the original CM are cited directly from [28]. On this dataset, our proposed SIRMF method obtains slight improvement in comparison to the original CM (i.e., 1.89° vs 2.0° in terms of median AE, 6% improvement). Furthermore, our proposed model also obtains quite competitive performance compared with other state-of-the-art learning-based methods (note that on this dataset, Barron's result [58] is quite beyond the performance of other learning-based methods).

Figure 5 displays two exemplar images. When the proposed model correctly selects the training images, both the test image and the selected training images are quite similar in both the illuminant and scene content (e.g., for the case #1, both the test and the selected training images are from the indoor environments and with the similar color-bias). However, when the proposed model wrongly selects the training images, both the test image and the selected images are quite different in both the illuminant and scene (as indicated by the case #2). We also list the training images selected by SIFT, which are mostly similar in the scene contents but with quite different

Fig. 5. The selected training images returned by the proposed model using the proposed feature $F^{'}$ and the SIFT on the Gehler-Shi dataset. The numbers reported in the second column are the angular errors of the test images corrected using the matrix derived from the selected images.

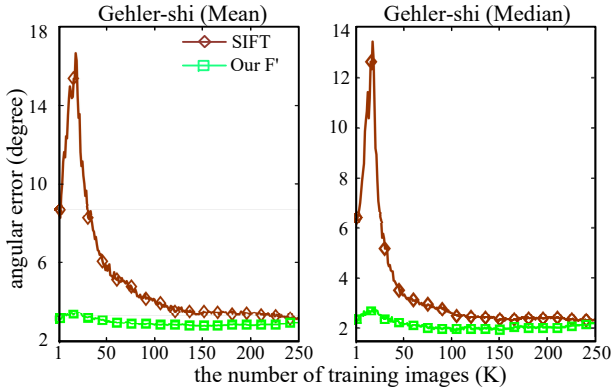

Fig. 6. The influence of the parameter K on the mean and median angular error on the Gehler-Shi dataset when using the SIFT feature and our proposed $F^{'}$ feature.

illuminants (e.g., the selected images include both the indoor and outdoor environments).

To further validate that our proposed feature $F^{'}$ is useful to find more suitable training images to improve the performance of the original CM, we tried to replace our proposed feature $F^{'}$ with SIFT and predict the illuminant estimate by the original CM in our framework. We ran this new version of our model on the Gehler-Shi dataset to evaluate the illuminant estimation performance under the condition of varying parameter K. In Fig. 6, the brown line indicates the results of the proposed model using the SIFT to select the training images for illuminant estimation. The model based on SIFT finds wrong images at the beginning and thus leads to very high angular error. With the increasing of parameter K, more similar images are selected into the training set, which results in a continuous decreasing of the angular error. However, our proposed method using the feature $F^{'}$ can always correctly select the suitable images no matter in the conditions of small K (e.g., $K = 1$) or large K values (e.g., $K = 100$), and thus always obtain better performance for illuminant estimation than using the SIFT. In the experiments, we found that SIFT often returns

Table 3. **AE and RAE of Various Methods on the Gehler-Shi Dataset.**

Methods	AE		RAE	
	Median	Mean	Median	Mean
WP [8]	5.68°	7.55°	6.5°	8.1°
GW [7]	6.28°	6.36°	6.8°	7.0°
NIS [37]	3.13°	4.19°	3.5°	4.8°
PBG [54]	2.3°	4.2°	2.7°	4.8°
EB [30]	2.3°	3.1°	2.6°	3.4°
CNN fine-tuned [19]	1.98°	2.63°	–	–
CNN+SVR [56]	1.44°	2.36°	–	–
SVRC_R [57]	1.97°	2.36°	–	–
Cheng et al. [21]	1.65°	2.42°	–	–
Barron [58]	1.22°	1.95°	–	–
AlexNet-$FC4$ [36]	1.11°	1.77°	–	–
Fast Fourier [59]	**0.86**°	**1.61**°	–	–
CM(19 edge) [28]	2.0°	2.8°	–	–
SIRMF(auto, K=86,107,84)	1.97°	2.83°	2.17°	3.33°
SIRMF(K=100)	1.89°	2.80°	**2.11**°	**3.29**°

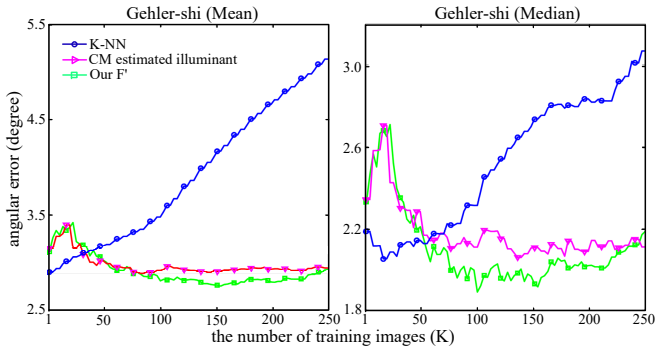

Fig. 7. The influence of the parameter K on the mean and median angular errors on the Gehler-Shi dataset when selecting suitable training images with the K-NN, CM-estimated illuminant and our proposed $F^{'}$ feature.

the images with similar scene contents as done by most of learning-based methods [30, 37–39]. In contrast, our proposed feature $F^{'}$ basically returns the images with similar lighting conditions (e.g., the # 1 first row in Fig. 5) that is quite helpful of learning an optimal correction matrix for each test image.

This conclusion is further grounded by Fig. 7, which shows the angular errors of mean and median by respectively using the CM-estimated illuminant and our proposed feature $F^{'}$ to select the similar images to re-train the correction matrix. The global tendency of the curves obtained by CM illuminant estimation and our $F^{'}$ is quite similar, which indicates that our $F^{'}$ indeed selects the suitable images with similar illuminants. Moreover, in comparison to the results by using

Optics EXPRESS

our feature F' (green line), the results based on the selection using CM illuminant estimation seems worse in terms of both mean and median AEs. We got similar observation on the SFU lab dataset (not shown here). In short, in comparison to the SIFT and CM estimated illumiants, our proposed feature F' is more useful to select the suitable images for re-training the correction matrix and thus further improve the performance of the original CM.

To further prove the effectiveness of the proposed F' feature, we also tested a weighted version of the K-Nearest Neighbors (K-NN) to select the illumination of training images as the illuminant estimation of each test image. The weight of each image was determined by the cosine distance between the F' features of the test image and the training images. We simply averaged the illumination of the selected K nearest illumination of training images as a baseline for comparison. The K-NN based mechanism directly picks the ground truth illuminant of training images according to the cosine distance measure, which is quite similar to the way used by the Exemplar-based color constancy [30]. We can see from the blue curves in Fig. 7 that the angular error of K-NN based mechanism is gradually increasing with the number of the selected K ground truth illuminant and its overall performance is worse than the mechanism based on our proposed F'. One main reason is that K-NN based selection mechanism simply averages the illumination of the selected images without any filtering mechanisms.

4.3. Grey ball dataset

Grey ball dataset contains 11346 images, which were mainly extracted from the video captured under 15 different locations. Due to there is very strong redundancy in the original dataset [4, 10, 28], Bianco et al. removed the correlation among images by video analysis and the final dataset totally contains 1135 images, which were splitted into 340 images for training and 795 images for test. It should be noted that the results of other methods listed in Table 4 were mostly based on the evaluation on the whole dataset [4] or a subset that just contains 150 images [28]. Table 4 shows that our proposed method performs best among all the methods evaluated here and an improvement of around 0.5 degree in the median of AE over the original CM.

4.4. NUS dataset

NUS dataset is one of the quite recent CC dataset, which is composed of 1736 high quality images collected by eight different commercial cameras [24]. To the best of our knowledge, the

Table 4. **AE and RAE of Various Methods on the Grey Ball Dataset.**

Methods	AE		RAE	
	Median	Mean	Median	Mean
WP [8]	6.30°	7.72°	5.5°	7.1°
GW [7]	5.68°	6.49°	7.6°	8.7°
GE2 [10]	5.08°	5.70°	5.0°	6.5°
NIS [37]	4.39°	5.14°	4.3°	5.5°
PBG [54]	5.8°	7.1°	5.9°	7.5°
Wu et al. [42]	2.90°	4.19°	–	–
EB [30]	3.4°	4.4°	3.7°	4.8°
CM(9 edge) [28]	3.3°	–	–	–
SIRMF(auto, K=110)	2.82°	4.56°	3.15°	5.02°
SIRMF(K=84)	**2.76°**	**4.46°**	**2.98°**	**4.89°**

Optics EXPRESS

Table 5. **AE of Various Methods on the NUS Dataset.**

Methods	WP	GW	GE2	NIS	Cheng	CM	CM(9 edge)*	Our(auto)	Our(best K)
Dataset	Median								
Canon1Ds	6.19°	4.15°	2.44°	3.04°	**1.57°**	1.98°	2.13°	1.82°	1.73°
Canon600D	12.44°	2.88°	2.29°	2.46°	**1.62°**	1.85°	1.74°	1.66°	1.63°
FujiXM1	10.59°	3.30°	2.00°	2.96°	**1.58°**	2.11°	1.96°	1.91°	1.66°
NikonD5200	11.67°	3.39°	2.19°	2.40°	1.65°	2.04°	1.84°	1.69°	**1.62°**
OlympEPL6	9.50°	2.58°	2.18°	2.17°	**1.41°**	1.84°	1.73°	1.61°	1.52°
LumixGX1	18.00°	3.06°	2.04°	2.28°	1.61°	1.77°	1.70°	1.53°	**1.48°**
SamNX2000	12.99°	3.00°	2.32°	2.77°	1.78°	1.85°	1.95°	1.81°	**1.64°**
SonyA57	7.44°	3.46°	2.70°	2.88°	**1.51°**	2.05°	1.83°	1.63°	**1.51°**
Dataset	Mean								
Canon1Ds	7.99°	5.16°	3.47°	4.18°	**2.26°**	2.94°	2.70°	2.57°	2.53°
Canon600D	10.96°	3.89°	3.21°	3.43°	**2.43°**	2.76°	2.52°	2.44°	2.45°
FujiXM1	10.20°	4.16°	3.12°	4.05°	**2.45°**	3.23°	2.69°	2.54°	2.64°
NikonD5200	11.64°	4.38°	3.47°	4.10°	**2.51°**	3.46°	2.80°	2.73°	2.69°
OlympEPL6	9.78°	3.44°	2.84°	3.22°	**2.15°**	2.95°	2.60°	2.53°	2.51°
LumixGX1	13.41°	3.82°	2.99°	3.70°	2.36°	3.10°	2.38°	2.28°	**2.23°**
SamNX2000	11.97°	3.90°	3.18°	3.66°	2.53°	2.74°	2.60°	2.53°	**2.37°**
SonyA57	9.91°	4.59°	3.36°	3.45°	**2.18°**	2.95°	2.28°	2.21°	**2.18°**

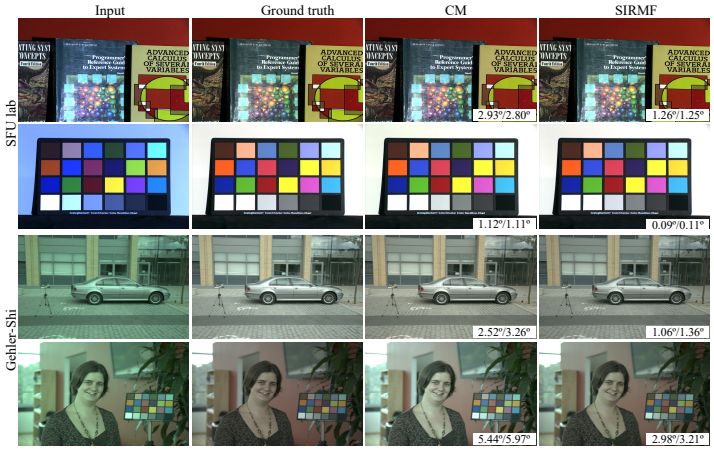

Fig. 8. Four images of the SFU lab (top two) and Gehler-Shi (bottom two) dataset corrected by the original CM and our proposed method. The numbers in the lower right corner show the angular errors and reproduction errors.

eight cameras used in this dataset have nearly similar camera spectral sensitivity curves and each camera contains almost the same scenes as other cameras. On this dataset, we found that the 9-edge moment based CM (marked as CM (9 edge) with asterisk in Tables 5 and 6) performs better than the 19-edge moment based CM, so we utilized the 9-edge moment based CM to evaluate our method. For better comparison, we also list the results of the CM implemented by [21].

Tables 5 and 6 show that our SIRMF model obtains at least 5% improvement in mean angular error and 10% improvement in median angular error compared with the original CM (our implementation), which are quite remarkable considering that the error obtained by the original CM is quite low. In comparison to the more recent regression-based CC algorithm proposed

Optics EXPRESS

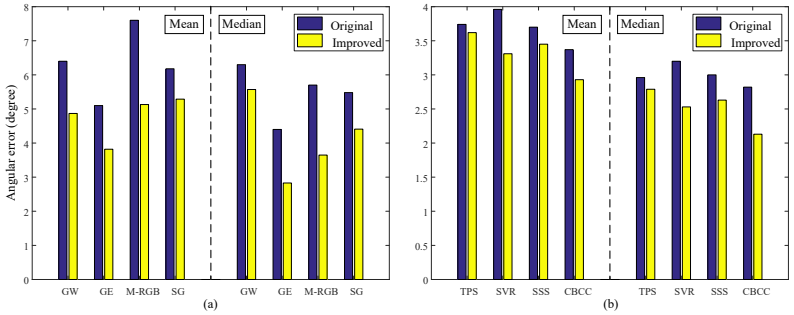

Fig. 9. Improvements of more methods on the Gehler-Shi dataset, (a) using gray world (GW) [7], gray edge (GE, 2nd-order) [10], Max-RGB (M-RGB) [8], shades of gray (SG, norm=4, sigma=4) [9] to replace CM to provide the initial illuminant. (b) using thin-plate spline interpolation (TPS) [60], support vector regression (SVR, 3D) [57,61], spatio-spectral statistics (SSS, ML Estimate) [55], committee-based CC (CBCC, edge moment) [47] to replace the original CM as the learning model and provide the initial illuminant.

in [21], our proposed model also obtains competitive performance in total, and even obtains better performance on several subsets. For example, on the sets of camera LumixGX1 and SamNX2000, the median AE of our method and Cheng's method are respectively SIRMF (1.48^o) vs Cheng (1.61^o) and SIRMF (1.64^o) vs Cheng (1.78^o). Interestingly, since our strategy is to try to pick the images with similar illuminants and Cheng et al. [21] also gathered the images with similar illuminants using regression tree based on four illuminant estimates, the good performance of [21] also proves the effectiveness of illuminant related feature for illuminant estimation. Our $F^{'}$ is a mathematically designed feature closely related to the illuminant, which could also be used to replace the four simple features proposed in [21] . In short, the main difference is that we use the proposed $F^{'}$ feature as a gating mechanism to filter out the unrelated training images when training a regression-based CC. In contrast, Cheng et al. developed an effective regression tree-based CC using four hand-crafted features. In other words, Cheng's method [21] is inherently similar to the original CM [28].

We also compared with two deep learning based methods in the Table 7, i.e., the fast Fourier CC (FFCC) algorithm [59] and the FC4 algorithm [36] on the NUS dataset. These two methods reported their performances for the whole NUS dataset by computing the geometric means of the measures of the eight subsets. Following their way, we also reported the performance of our method on the whole NUS dataset by computing the geometric means of the measures of the eight subsets listed in Table 5. We can see from Table 7 that the two variations of our model can provide quite acceptable performance in comparison to these two methods.

To further investigate the generalization ability of our approach, we also conducted the inter-dataset-based evaluation on the NUS dataset. That is, we trained the learning-based model on one subset (e.g., Canon1Ds) and tested it on another subset (e.g., SamNX2000). Since the NUS dataset was captured by different cameras (each camera has its own intrinsic properties) under various exposure conditions [24], training a CC model on one camera and then testing it on another camera is a much more difficult task than the previous experiments. To the best of our knowledge, there is no previous CC work considering such type of inter-dataset evaluation on the NUS dataset except [2]. Table 8 lists four inter-dataset evaluations on the NUS dataset. We can observe that our proposed method can greatly increase the performance of the original CM with near 10% when conducting such a quite difficult inter-dataset evaluation. It should noted that the sensors of each subset in the NUS dataset are different, but the differences are not too large [2], which may help our proposed strategy get quite accurate $F^{'}$ feature and hence the quite accurate

Optics EXPRESS

Table 6. **RAE of Various Methods on the NUS Dataset.**

Methods	CM(9 edge)*	Our(auto)	Our(best K)	CM(9 edge)*	Our(auto)	Our(best K)
Dataset	Median			Mean		
Canon1Ds	2.57°	**2.08°**	2.20°	3.65°	**3.37°**	3.40°
Canon600D	2.30°	2.14°	**2.13°**	3.38°	**3.19°**	3.33°
FujiXM1	2.60°	2.54°	**2.29°**	3.76°	**3.58°**	3.78°
NikonD5200	2.66°	2.30°	**2.27°**	4.04°	3.91°	**3.86°**
OlympEPL6	2.27°	2.08°	**2.05°**	3.59°	3.45°	**3.42°**
LumixGX1	2.30°	2.05°	**1.99°**	3.34°	3.21°	**3.04°**
SamNX2000	2.60°	2.45°	**2.26°**	3.55°	3.42°	**3.19°**
SonyA57	2.32°	2.19°	**2.05°**	3.14°	2.99°	**2.95°**

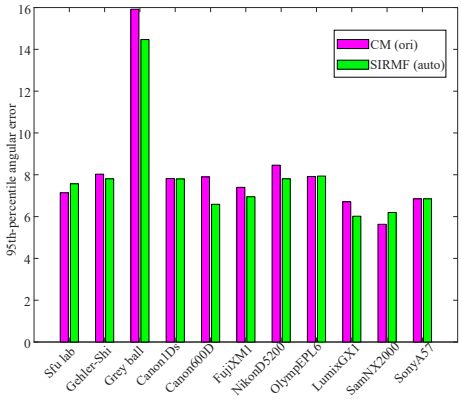

Fig. 10. Comparison of the 95th-percentile AE values of the original CM (Ori) and Our SIRMF(auto) for each dataset.

illuminant estimation. Table 8 indicates that if the different sensors are not sufficiently diverse, our method may capture something fundamental about the illuminant estimation process that may suffice for general applicability across different image capturing configurations. We finally trained our model on the Gehler-Shi dataset and then tested it on the NUS dataset or vice versa. We can clearly see from Table 9 that our SIRMF(auto) performs well for this inter-dataset based cross-validation, where both sensors and illuminants appeared in Gehler-Shi dataset are not at all represented by the NUS dataset.

Figure 8 lists the examples of the color constancy results by the proposed method on the SFU lab and Gehler-Shi datasets. These pictures show that compared with the original CM, our proposed method clearly improves the results in both the visual appearance and the quantitative measures of angular error and reproduction error.

Table 7. **Results of Two Deep Learning Based Methods [36, 59] and our SIRMF on the NUS Dataset.**

Methods	FC4	FFCC	SIRMF (auto)	SIRMF (best k)
Median AE	1.53°	1.31°	1.60°	1.70°
Mean AE	2.23°	1.99°	2.44°	2.47°

Optics EXPRESS

Table 8. **Inter-dataset Based Evaluation on the NUS Dataset.**

Methods	CM(9 edge)*	SIRMF(auto)
Dataset	Median	
Canon1Ds→Canon600D	2.01°	**1.89°**
Canon600D→Canon1Ds	2.29°	**2.11°**
Canon1Ds→SamNX2000	4.23°	**4.03°**
SamNX2000→Canon1Ds	3.69°	**3.33°**
Dataset	Mean	
Canon1Ds→Canon600D	2.54°	**2.43°**
Canon600D→Canon1Ds	2.86°	**2.67°**
Canon1Ds→SamNX2000	4.82°	**4.80°**
SamNX2000→Canon1Ds	4.26°	**4.03°**

Table 9. **Training SIRMF on the Gehler-Shi Dataset and Testing it on the NUS Dataset.**

Method	SIRMF(auto)	
Dataset	Mean	Median
Gehler-Shi→Canon1Ds	5.33°	4.85°
Gehler-Shi→Canon600D	5.57°	5.32°
Gehler-Shi→FujiXM1	5.63°	4.96°
Canon1Ds→Gehler-Shi	5.38°	4.65°
Canon600D→Gehler-Shi	5.67°	5.08°
FujiXM1→Gehler-Shi	6.08°	5.62°

Considering the long-tail nature of the angular error distribution, we finally present the 95th-percentile AE values for all the datasets. Figure 10 shows the results of the original CM and the improved results using our SIRMF. We find that our strategy can also improve the worst performance of original CM on most of the datasets.

4.5. Generalization of the proposed framework

Although the above methodology description and experiments focus on improving the regression-based CC method of CM by using the proposed automatic training set selection strategy, we believe that the proposed method is general and applicable to many other regression or learning based CC methods. For example, we used the CC method of CM in this work to first get an illuminant estimation for the calculation of F' of the test image, we can replace CM with other CC methods (e.g., low-level based methods) for illuminant estimation at this step. To validate this point, we selected four statistic-based methods (i.e., GW [7], GE [10], Max-RGB [8] and SG [9]) to replace the method of CM to provide an initial illuminant for the test image in the stage of selecting suitable training set, while keeping all the operations of other steps the same as the original framework shown in Fig. 1. The results shown in Fig. 9(a) indicate that the

Table 10. **Results of the Best Low-level CC and the Automatic Algorithm Selection Based on the F' Measure for the SFU Lab Dataset and the Gehler-Shi Dataset.**

Methods	Low-level CC		F' measure	
Dataset	Median	Mean	Median	Mean
SFU lab (GE2)	2.74°	5.23°	2.68°	4.70°
Gehler-Shi (SG)	4.01°	4.93°	3.45°	4.54°

proposed framework can improve the original statistic-based methods over than 20% in terms of the mean and median angular errors. In general, the more accurate illuminant estimation is provided by a CC method at this step, the more accurate $F^{'}$ we can get, and a better refined illuminant estimation can be finally obtained.

As another more convincing attempt to validate the generalization ability of the proposed framework, we selected four representative learning-based methods (i) to replace the method of CM to provide an initial illuminant for the test image in the stage of selecting suitable training set and (ii) to replace the method of CM in the stage of training and testing in the original framework shown in Fig. 1. As indicated in Fig. 9(b), we obtain over than 10% improvements in terms of the mean and median angular errors compared to the original learning-based methods listed here.

We finally investigated the function of proposed $F^{'}$ measure. Specifically, we firstly selected five low-level based CC approaches including white patch (WP), grey world (GW), 1st-order Grey-Edge (GE1), 2nd-order Grey-Edge (GE2), and shades of grey (SG). Then, based on the $F^{'}$ measure for a given test image, we chose the training images as described in the previous section. With the selected training images, the corresponding CC algorithm from the above list that performs best on the selected training images is selected and finally applied to the test image. We evaluated how the proposed $F^{'}$ measure improves the low-level based CC algorithm that performs well on each dataset (e.g., GE2 performs best on the SFU lab dataset and SG performs best on the Gehler-Shi dataset). We can see from Table 10 that the proposed $F^{'}$ measure could effectively improve the performance of existing low-level based methods in terms of both mean and median AEs on these two datasets.

For more general applications such as an illuminant (e.g., there are no "ground truth" for natural scene images) that is not at all represented by the dataset, we may first create a minimum training set that contains a variety of typical lighting scenes, and then pre-learn a correction matrix for each typical illuminant. When a new test image comes, the model can choose the most suitable correction matrix according to the calculated moment $F^{'}$. In addition, the moment vectors of all the training images can be previously computed, which can be directly used for the computation of moment distance when no suitable "ground truth" is available.

4.6. *Deviation of automatically selected K from the optimal and the possible explanation*

Our K is automatically estimated in an average sense for the whole dataset. To explore the difference between the automatically selected K and the optimal K required per test image, we first ran all the possible K values for each test image by exhaustive searching and defined $K_{opt}(id)$ as the optimal K value that produces the lowest AE, where id denotes a test image, indicating that different test images may require different K_{opt}. Then, we defined the deviation ratio as $(K_{opt}(id)-K)/N$, where N is the number of total training images in a dataset.

The scatter plots of deviation ratios for the three representative datasets are shown in Fig. 11. We computed the linear average of the deviation ratios over all the test images for each dataset in Fig. 11, and we obtained 0.2416, -0.0461 and 0.0229 for the SFU Lab, Gehler-Shi and NUS-FujiXM1 datasets, respectively, as illustrated by the dashed horizontal lines in Fig. 11. It is clear that except for the SFU Lab dataset, the values of these average deviation ratios are quite close to zero, no matter the average deviation ratio is lower (Gehler-Shi) or higher (NUS-FujiXM1) than zero. This indicates that in the sense of linear average, the automatically determined values of global K are quite close the values of optimal K_{opt} for all the test images contained in the datasets. In other words, from the point of view of average deviation ratio, the bias between the global K and the individual optimal K_{opt} values of all the test images can be almost neglected, as illustrated by the quite close distances between the solid and the dashed horizontal lines for the Gehler-Shi and NUS-FujiXM1 datasets.

That is to say, for the Gehler-Shi and NUS-FujiXM1 datasets, although the deviation ratios of

Optics EXPRESS

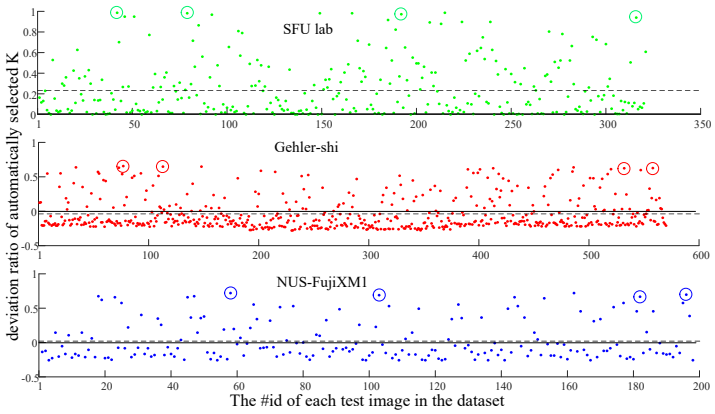

Fig. 11. The distribution of the deviation ratio of the automatically selected K for the test images of the three datasets. Each dot corresponds to the deviation ratio of the optimal K for one test image (i.e., $K_{opt}(id)$). On each panel, the solid horizontal line denotes the deviation ratio of zero, and the dashed line denotes the average of the deviation ratios of all the test images. The average deviation ratios are respectively 0.2416, -0.0461 and 0.0229 for the SFU Lab, Gehler-Shi and NUS-FujiXM1 datasets. The quite close distances between the solid and the dashed horizontal lines for the Gehler-Shi and NUS-FujiXM1 datasets indicate that in an average sense, the bias between the global K and the individual optimal K_{opt} values of all the test images can be almost neglected.

all the test images are not distributed evenly around zero (i.e., the majority of the dots have small deviation ratios and are distributed below (but close to) zero, and the minority of the dots have quite large deviation ratios and are distributed above (but far from) zero), they are concentrated around zero in an average sense.

It can be easily imagined that if all the individual optimal numbers $K_{opt}(id)$ required by different test images are evenly distributed, the "optimal parameter K" should make all the deviation ratios (computed with deviation ratio=$(K_{opt}(id)-K)/N$) distributed evenly around zero. However, it seems that the $K_{opt}(id)$ values for different test images are NOT evenly distributed. As illustrated in Fig. 11, the uneven distribution of deviation ratios holds true for all the three datasets, from which we can easily understand that the distribution of all $K_{opt}(id)$ is also uneven, with the majority of test images having small $K_{opt}(id)$ and the minority having quite large $K_{opt}(id)$. With such biased distribution of $K_{opt}(id)$, we can never obtain an even distribution of deviation ratios no matter what a global K value is selected.

To get an insight understanding of the deviation distribution shown in Fig. 11, we pick out four examples of the test images with the quite high deviation ratios (marked with the circles in Fig. 11) for each dataset and list them in the Fig. 12. We found that such test images always contain quite simple scenes (Fig. 12). We speculated that more training images would be required for these simple scenes in order to fully learn and code the implicit (e.g., high-order) features of these scenes that can discriminate them from others, which is the reason resulting in the much higher deviation ratios for these scenes.

We have also conducted a new experiment on the Gehler-shi dataset to check whether or not our selection method can pick a global K value that makes the deviation ratios distributed evenly around zero for a dataset of relatively less diversity. As demonstrated above, the images with quite high deviation ratios are out of the ordinary. Therefore, we first excluded 43 images with the deviation ratios above a threshold of 0.4 from the whole Gehler-shi dataset, and the subset composed by the remaining 525 images is of course of relatively less diversity. Then, the value of global K was automatically determined for this subset, and the deviation ratios were computed

Optics EXPRESS

Fig. 12. Four example images from each dataset with high deviation ratios (marked with the circles in Fig. 11).

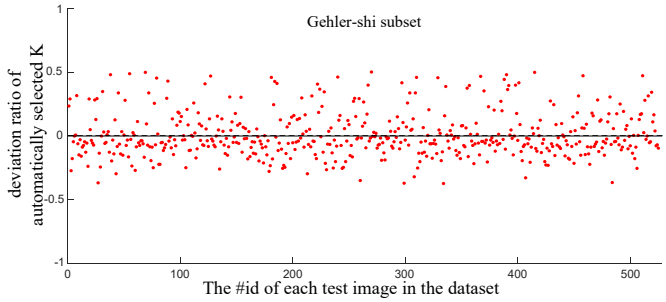

Fig. 13. The distribution of the deviation ratios of the automatically selected K for the test images of the Gehler-Shi subset after excluding the images with the high deviation ratios above a threshold of 0.4 from the whole dataset. The solid horizontal line denotes the deviation ratio of zero, and the dashed line denotes the average deviation ratios, which is 0.0062.

for the 525 images, as shown in Fig. 13. From Fig. 13, we can clearly see that the deviation ratios for this subset are distributed more evenly around the zero in comparison to that shown in Fig. 11. In addition, the average of the deviation ratios in Fig. 13 is 0.0062, which is significantly lower than -0.0461 in Fig. 11 in terms of absolute value. The results of this experiment further support that our selection method could automatically find a good training set size in an average sense that is suitable for most of the test images of common diversity.

4.7. Perceptual meaningful measures

Considering that what is finally evaluating the images is human visual system, we further calculated respectively the errors in the CIELuv space and CIELab space. In the experiments, we calculated the Euclidean distance between two vectors in the CIELuv space (e.g., the illuminant ground truth and the estimated illuminant). This Euclidean distance shown in Table 11 is represented as CIELuv DE. We can see that our proposed method (SIRMF(auto)) significantly improves the performance of CM in terms of this perceptually uniform space. Figure 14 also shows that there exists significantly positive correlation between the AE in RGB space and the DE in CIELUV space. Moreover, we further used the Delta E2000 [62] to verify whether our

Table 11. **Perceptually Meaningful Measures Evaluated on Two Datasets.**

Methods	CM		SIRMF(auto)	
Dataset	CIELuv DE	Delta E2000	CIELuv DE	Delta E2000
SFU lab	5.66	2.56	4.86 (14%)	2.05 (20%)
Gehler-Shi	5.67	2.94	5.23 (7%)	2.76 (6%)

Optics EXPRESS

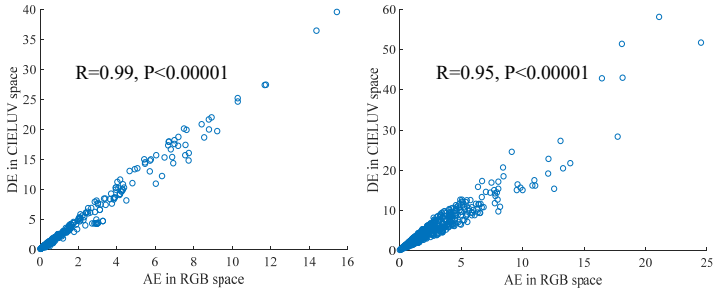

Fig. 14. Correlation between the AE in RGB space and the DE in CIELUV space for the SFU lab dataset (Left) and the Gehler-Shi dataset (Right).

proposed SIRMF(auto) substantially improves the existing methods based on the perceptual measures. Results of Delta E2000 for our SIRMF(auto) and CM on two datasets are also shown in Table 11. We can see that our SIRMF(auto) significantly improves the results of CM in terms of Delta E2000.

5. Conclusion and discussion

We proposed a method to improve the regression-learning-based CC methods by selecting suitable training images for each test image. One contribution is to introduce a novel feature F' to improve the existing image edge (IE) based methods. The feature F' is directly derived from the regression-based equation and the image formation model. Experimental results show that using F' is more effective than directly using the estimated illuminant and other features (e.g., SIFT) to select suitable training images. In particular, our proposal using this simple feature F' with only 9 dimensions can automatically find a good training set size in an average sense that is suitable for most of the test images. Moreover, our proposal can obtain quite competitive performance compared to many state-of-the-art approaches and also significantly improve the performance of several regression learning-based CC algorithms using such a simple idea and simple feature.

Funding

National Natural Science Foundation of China (NSFC) (61806134); Fundamental Research Funds for the Central Universities (YJ201751); Sichuan Province Science and Technology Support Project (2017SZDZX0019).

Acknowledgments

The authors would like to thank Dr. Xiaochuan Chen for sharing their code of CM.

References

1. X. Yang, X. Jin, and J. Zhang, "Improved single-illumination estimation accuracy via redefining the illuminant-invariant descriptor and the grey pixels," Opt. express **26**, 29055–29067 (2018).
2. S. B. Gao, M. Zhang, C. Y. Li, and Y. J. Li, "Improving color constancy by discounting the variation of camera spectral sensitivity," J. Opt. Soc. Am. A **34**, 1448–1462 (2017).
3. D. An, J. Suo, H. Wang, and Q. Dai, "Illumination estimation from specular highlight in a multi-spectral image," Opt. express **23**, 17008–17023 (2015).
4. A. Gijsenij, T. Gevers, and J. Van De Weijer, "Computational color constancy: Survey and experiments," IEEE Transactions on Image Process. **20**, 2475–2489 (2011).
5. Sivalogeswaran, Ratnasingam and Steve, Collins and Javier, Hernández-Andrés, "Optimum sensors for color constancy in scenes illuminated by daylight," J. Opt. Soc. Am. A **27**, 2198–2207 (2010).
6. Sivalogeswaran, Ratnasingam and Steve, Collins and Javier, Hernández-Andrés, "Extending "color constancy" outside the visible region," J. Opt. Soc. Am. A **28**, 541–547 (2011).
7. G. Buchsbaum, "A spatial processor model for object colour perception," J. Frankl. institute **310**, 1–26 (1980).

Optics EXPRESS

8. E. H. Land, *The retinex theory of color vision* (Citeseer, 1977).
9. G. D. Finlayson and E. Trezzi, "Shades of gray and colour constancy," in *Color and Imaging Conference,* vol. 2004 (Society for Imaging Science and Technology, 2004), pp. 37–41.
10. J. Van De Weijer, T. Gevers, and A. Gijsenij, "Edge-based color constancy," IEEE Transactions on Image Process. **16**, 2207–2214 (2007).
11. A. Gijsenij, T. Gevers, and J. Van De Weijer, "Improving color constancy by photometric edge weighting," IEEE Transactions on Pattern Analysis Mach. Intell. **34**, 918–929 (2012).
12. H. R. V. Joze, M. S. Drew, G. D. Finlayson, and P. A. T. Rey, "The role of bright pixels in illumination estimation," in *Color and Imaging Conference,* vol. 2012 (Society for Imaging Science and Technology, 2012), pp. 41–46.
13. K.-F. Yang, S.-B. Gao, and Y.-J. Li, "Efficient illuminant estimation for color constancy using grey pixels," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* (2015), pp. 2254–2263.
14. S. B. Gao, W. W. Han, K. F. Yang, C. Y. Li, and Y. J. Li, "Efficient color constancy with local surface reflectance statistics," in *European Conference on Computer Vision,* (Springer, 2014), pp. 158–173.
15. X. S. Zhang, S. B. Gao, R. X. Li, X. Y. Du, C. Y. Li, and Y. J. Li, "A retinal mechanism inspired color constancy model," IEEE Transactions on Image Process. **25**, 1219–1232 (2016).
16. S. B. Gao, K. F. Yang, C. Y. Li, and Y. J. Li, "Color constancy using double-opponency," IEEE Transactions on Pattern Analysis Mach. Intell. **37**, 1973–1985 (2015).
17. S. B. Gao, K. F. Yang, C. Y. Li, and Y. J. Li, "A color constancy model with double-opponency mechanisms," in *Proceedings of the IEEE International Conference on Computer Vision,* (2013), pp. 929–936.
18. K. A. Smet, Q. Zhai, M. R. Luo, and P. Hanselaer, "Study of chromatic adaptation using memory color matches, part ii: colored illuminants," Opt. express **25**, 8350–8365 (2017).
19. 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.
20. S. W. Oh and S. J. Kim, "Approaching the computational color constancy as a classification problem through deep learning," Pattern Recognit. **61**, 405–416 (2017).
21. D. Cheng, B. Price, S. Cohen, and M. S. Brown, "Effective learning-based illuminant estimation using simple features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* (2015), pp. 1000–1008.
22. W. Shi, C. C. Loy, and X. Tang, "Deep specialized network for illuminant estimation," in *European Conference on Computer Vision,* (Springer, 2016), pp. 371–387.
23. D. A. Forsyth, "A novel algorithm for color constancy," Int. J. Comput. Vis. **5**, 5–35 (1990).
24. 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).
25. K. Barnard, L. Martin, B. Funt, and A. Coath, "A data set for color research," Color. Res. & Appl. **27**, 147–151 (2002).
26. L. Shi and B. Funt, "Re-processed version of the gehler color constancy dataset of 568 images," http://www.cs.sfu.ca/colour/data/. .
27. F. Ciurea and B. Funt, "A large image database for color constancy research," in *Color and Imaging Conference,* vol. 2003 (Society for Imaging Science and Technology, 2003), pp. 160–164.
28. G. D. Finlayson, "Corrected-moment illuminant estimation," in *Proceedings of the IEEE International Conference on Computer Vision,* (2013), pp. 1904–1911.
29. Z. Lou, T. Gevers, N. Hu, and M. P. Lucassen, "Color constancy by deep learning," in *British Machine Vision Conference,* (2015), pp. 76–81.
30. H. R. V. Joze and M. S. Drew, "Exemplar-based color constancy and multiple illumination," IEEE Transactions on Pattern Analysis Mach. Intell. **36**, 860–873 (2014).
31. Marc Ebner and Johannes Hansen, "Depth map color constancy," Bio-Algorithms Med-Systems **9**, 167–177 (2013).
32. Marc Ebner, *Color Constancy*, Wiley-IS&T series in imaging science and technology (Wiley, 2007).
33. K. Barnard, G. Finlayson, and B. Funt, "Color Constancy for Scenes with Varying Illumination," Comput. Vis. Image Underst. **65**, 311 – 321 (1997).
34. S. Gao, Y. Ren, M. Zhang, and Y. Li, "Combining bottom-up and top-down visual mechanisms for color constancy under varying illumination," IEEE Transactions on Image Process. **28**, 4387–4400 (2019).
35. Y. Qian, K. Chen, J.-K. Kämäräinen, J. Nikkanen, and J. Matas, "Deep structured-output regression learning for computational color constancy," in *Pattern Recognition (ICPR), 2016 23rd International Conference on,* (IEEE, 2016), pp. 1899–1904.
36. 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.
37. A. Gijsenij and T. Gevers, "Color constancy using natural image statistics and scene semantics," IEEE Transactions on Pattern Analysis Mach. Intell. **33**, 687–698 (2011).
38. S. Bianco, G. Ciocca, C. Cusano, and R. Schettini, "Improving color constancy using indoor–outdoor image classification," IEEE Transactions on Image Process. **17**, 2381–2392 (2008).
39. J. Van de Weijer, C. Schmid, and J. Verbeek, "Using high-level visual information for color constancy," in *2007 IEEE 11th International Conference on Computer Vision,* (IEEE, 2007), pp. 1–8.
40. B. Li, W. Xiong, W. Hu, and H. Peng, "Illumination estimation based on bilayer sparse coding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* (2013), pp. 1423–1429.

Optics EXPRESS

41. M. Wu, J. Sun, J. Zhou, and G. Xue, "Color constancy based on texture pyramid matching and regularized local regression," J. Opt. Soc. Am. A **27**, 2097–2105 (2010).
42. M. Wu, K. Luo, J. Dang, and J. Zhou, "Edge-moment-based color constancy using illumination-coherent regularized regression," J. Opt. Soc. Am. A **32**, 1707–1716 (2015).
43. B. A. Wandell, "The synthesis and analysis of color images," IEEE Transactions on Pattern Analysis Mach. Intell. **1**, 2–13 (1987).
44. G. D. Finlayson, R. Zakizadeh, and A. Gijsenij, "The reproduction angular error for evaluating the performance of illuminant estimation algorithms," IEEE Transactions on Pattern Analysis Mach. Intell. **39**, 1482–1488 (2017).
45. H. Y. Chong, S. J. Gortler, and T. Zickler, "The von kries hypothesis and a basis for color constancy," in *2007 IEEE 11th International Conference on Computer Vision,* (2007), pp. 1–8.
46. G. D. Finlayson, M. Mackiewicz, and A. Hurlbert, "Color correction using root-polynomial regression," IEEE Transactions on Image Process. **24**, 1460–1470 (2015).
47. V. C. Cardei and B. Funt, "Committee-based color constancy," in *Color and Imaging Conference,* vol. 1999 (Society for Imaging Science and Technology, 1999), pp. 311–313.
48. G. D. Finlayson and S. D. Hordley, "Reevaluation of color constancy algorithm performance," J. Opt. Soc. Am. A **23**, 1008–1020 (2006).
49. X. Chen, M. S. Drew, Z.-N. Li, and G. D. Finlayson, "Extended corrected-moments illumination estimation," Electron. Imaging **2016**, 1–8 (2016).
50. P. V. Gehler, C. Rother, A. Blake, T. Minka, and T. Sharp, "Bayesian color constancy revisited," in *2008 IEEE Conference on Computer Vision and Pattern Recognition,* (2008), pp. 1–8.
51. http://www.kyb.mpg.de/bs/people/pgehler/colour/index.html. .
52. S. Bianco, G. Ciocca, C. Cusano, and R. Schettini, "Automatic color constancy algorithm selection and combination," Pattern Recognit. **43**, 695–705 (2010).
53. A. Gijsenij, "Color constancy: research website on illuminate estimation," accessed from http://www.colorconstancy.com/. .
54. A. Gijsenij, T. Gevers, and J. Van De Weijer, "Generalized gamut mapping using image derivative structures for color constancy," Int. J. Comput. Vis. **86**, 127–139 (2010).
55. A. Chakrabarti, K. Hirakawa, and T. Zickler, "Color constancy with spatio-spectral statistics," IEEE Transactions on Pattern Analysis Mach. Intell. **34**, 1509–1519 (2012).
56. S. Bianco, C. Cusano, and R. Schettini, "Single and multiple illuminant estimation using convolutional neural networks," IEEE Transactions on Image Process. (2017).
57. B. Li, W. Xiong, W. Hu, and B. Funt, "Evaluating combinational illumination estimation methods on real-world images," IEEE Transactions on Image Process. **23**, 1194–1209 (2014).
58. J. T. Barron, "Convolutional color constancy," in *Proceedings of the IEEE International Conference on Computer Vision,* (2015), pp. 379–387.
59. J. T. Barron and Y.-T. Tsai, "Fast fourier color constancy," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, July,* (2017).
60. B. Funt, L. Shi, and W. Xiong, "Illumination estimation via thin-plate spline interpolation," J. Opt. Soc. Am. A **28**, 940 (2011).
61. W. Xiong and B. Funt, "Estimating illumination chromaticity via support vector regression," J. Imaging Sci. Technol. **50**, 47–52 (2006).
62. G. Sharma, W. Wu, and E. N. Dalal, "The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," Color. Res. & Appl. **30**, 21–30 (2005).