Bayesian Color Constancy Revisited

Peter Vincent Gehler
Max Planck Institute
Tübingen, Germany
pgehler@tuebingen.mpg.de

Carsten Rother, Andrew Blake, Tom Minka, Toby Sharp
Microsoft Research Cambridge
Cambridge, UK
{carrot,blake,minka,tsharp}@microsoft.com

Abstract

Computational color constancy is the task of estimating the true reflectances of visible surfaces in an image. In this paper we follow a line of research that assumes uniform illumination of a scene, and that the principal step in estimating reflectances is the estimation of the scene illuminant. We review recent approaches to illuminant estimation, firstly those based on formulae for normalisation of the reflectance distribution in an image — so-called grey-world algorithms, and those based on a Bayesian formulation of image formation.

In evaluating these previous approaches we introduce a new tool in the form of a database of 568 high-quality, indoor and outdoor images, accurately labelled with illuminant, and preserved in their raw form, free of correction or normalisation. This has enabled us to establish several properties experimentally. Firstly automatic selection of grey-world algorithms according to image properties is not nearly so effective as has been thought. Secondly, it is shown that Bayesian illuminant estimation is significantly improved by the improved accuracy of priors for illuminant and reflectance that are obtained from the new dataset.

1. Introduction

Color constancy is the tendency to perceive surface color consistently, despite variations in ambient illumination [11]. Most generally, illumination variations occur both within scenes, and from scene to scene, and theories such as the "Retinex" [12] have been devised to explain color constancy under such conditions. Here we address only the problem of variation from scene to scene, making the common assumption that illumination within a given scene is approximately uniform. Some theories seek to compute invariant descriptors of color in order to facilitate such tasks as object recognition and tracking *e.g.* [9]. Other theories address the problem of estimating the illuminant [8, 14, 3, 6, 1, 13, 16, 10, 15] and this allows constancy to be achieved by recoloring any given image under a standard illuminant.

Figure 1. Example results on an image from our new Color Checker Database. The upper image is taken with a Canon 1D in autowhitebalance mode. The lower image was corrected using the algorithm proposed in this paper.

In this paper we are concerned with illuminant estimation, and this has been tackled in several ways: by gamut mapping, by reflectance normalisation, and by Bayesian estimation. Each of the approaches models image intensity as a product of a uniform illuminant with a reflectance function over the visible scene. It is clear that the problem is underconstrained in principle, but this can be resolved by exploiting assumptions about the variability of scene reflectance.

In gamut mapping [8], the gamut of colors in a test image is remapped to the visible gamut under a standard illuminant, and that mapping constrains the illuminant. In normalisation approaches, the range of reflectances is normalised in each color channel, and the adjustment required for normalisation yields an estimate of the illuminant. For example, "scale-by-max" takes the maximum intensity in each color channel, and maps the resulting tristimulus vector to white reflectance [12]. This is a special case of the class of "grey-world" algorithms which map the mean reflectance of a test image, under the p-norm, to grey [7]. A variation of this is the "grey-edge" algorithm [16] which instead maps the mean contrast of edges to grey. Recently a statistical fusion of grey-world algorithms has used a classifier to try to select the optimal algorithm for a given image [10].

Bayesian approaches [14, 3, 13] model the variability of reflectance and of illuminant as random variables, and then estimate illuminant from the posterior distribution conditioned on image intensity data. Earlier attempts to do this did not outperform gamut mapping [1], but the use of non-parametric statistical models for reflectance, exploiting the insight of the "color by correlation" method [6] to capture the tendency of nearby pixels to be correlated, finally produced results of leading quality [13]. It is interesting to note that the grey-world algorithms are special cases of Bayesian estimation, so it is to be expected that Bayesian models should exist that outperform grey-world algorithms [13].

Earlier research on statistical approaches to illuminant estimation used synthetic data, but more recently the "grey-ball" set of real images [4], in which a grey sphere is included in each scene to allow the true illuminant to be calculated, has been used in several studies [10, 15]. This dataset was captured by video camera which has the advantage that a large number (more than 11000) of images can be collected. However there are some limitations. Firstly, nearby images in the video tend to be correlated, so care must be taken when sampling from the image-set to avoid correlation. Secondly, the camera has gamma correction built in and this makes precise photometric calibration more difficult. A further dataset [2] is available with labelled illuminant, but containing only 30 indoor scenes, under various illuminants, which is useful for testing purposes but not sufficiently diverse for training.

The paper reports on three new results. Firstly we have introduced a new dataset[1], captured using a high-quality digital SLR camera in RAW format, free of color correction. A copy of the "Macbeth" color chart is included in every scene and this allows us to compute accurate illumination labels for each image. A total of 568 images are available, both indoor (246) and outdoor (322). Secondly we examine the fusion algorithm [10] which has appeared to give performance better than any individual greyworld algorithm. Using the new database, rather than the grey-ball database with its high degree of correlation among images, we have shown that the fused algorithm is after all not significantly better than the best greyworld algorithm. Thirdly, we have revisited the Bayesian approach of Rosenberg et al. [13] using the new data-set. Where Rosenberg et al. [13] trained using illumination labels that were only estimated (by scale-by-max), the new dataset provides accurate illumination labels. This should make it possible to learn more precise priors for illumination and reflectance. Tests show that this leads to illuminant estimation for which the improvement in accuracy is statistically significant, at least for outdoor images. The newly trained Bayesian algorithm is shown also to perform significantly better than the grey-world algorithms, even when the greyworld algorithms are enhanced by inclusion of an illumination prior.

[1] available online together with the code at *research.microsoft.com/vision/cambridge*

2. A Bayesian Approach to Color Constancy

This section reviews the framework of [13]. Let $\mathbf{y}$ be an image pixel with three color channels (y_r, y_g, y_b) represented in a linear RGB space. The pixel value is assumed to the reflection of a single light source off of a Lambertian surface. The light has power (l_r, l_g, l_b) ranging from zero to infinity in the three color channels. The surface reflects some proportion of this light. We call this proportion the *reflectance* (x_r, x_g, x_b) and it ranges from zero to one in each color channel. Thus the model for an observed pixel is

$$y_c = l_c x_c, \qquad c = \{r, g, b\} \tag{1}$$

or with $\mathbf{L} = \text{diag}(l)$ $\mathbf{y} = \mathbf{L}\mathbf{x}$. An image is given as $\mathbf{Y} = (\mathbf{y}(1), \ldots, \mathbf{y}(N))$ with unknown reflectances $\mathbf{X} = (\mathbf{x}(1), \ldots, \mathbf{x}(N))$. We assume that the illumination and the reflectances are independent thus $p(x, l) = p(x)p(l)$. The illumination prior distribution $p(l)$ can be estimated from real world data and will be further described in Section 2.2. The most challenging part is to derive a model $p(x)$ for the reflectances in an image. The approach taken in [13] is to assume exchangeability of the reflectances. With this assumption one only needs to define probabilities over reflectance histograms $(n_1, \ldots, n_K)$ in an image where n_k are the number of reflectances in the kth bin of the histogram

$$p(\mathbf{X}) \propto f(n_1, \ldots, n_K) \tag{2}$$

The model used in [13] is

$$f(n_1, \ldots, n_K) = \prod_k m_k^{\nu_k} \tag{3}$$

$$\nu_k = n \frac{\text{clip}(n_k)}{\sum_s \text{clip}(n_s)} \tag{4}$$

$$\text{clip}(n_k) = \begin{cases} 0 & \text{if } n_k = 0 \\ 1 & \text{if } n_k > 0 \end{cases} \tag{5}$$

with m_k being the probability of a surface having a reflectance value in bin k. Practically bin clipping is very important. Consider an image with one domain surface which has everywhere the same reflectance, e.g. a big red wall (see Figure 2). Without clipping the red wall would considerably skew the reflectance distribution, however, clipping will balance that each surface in the image has more equal distribution. Additionally we experimented with a more moderate clipping function and replace ν_k in Equation (4) with

$$\nu_k = n\frac{\tanh(\lambda n_k)}{\sum_s \tanh(\lambda n_s)}, \tag{6}$$

where the parameter λ controls the shape of the function. For high values of λ almost all bin entries will be clipped whereas low values of λ correspond to a softer clipping behaviour.

The likelihood of the observed data for an illuminant $\mathbf{L}$ is

$$p(\mathbf{Y}|\mathbf{L}) = \int_{\mathbf{X}} \left(\prod_i p(\mathbf{y}(i)|\mathbf{L}, \mathbf{x}(i)) \right) p(\mathbf{X}) d\mathbf{X} \tag{7}$$

$$= |\mathbf{L}^{-1}|^n p(\mathbf{X} = \mathbf{L}^{-1}\mathbf{Y}) \tag{8}$$

and the posterior for $\mathbf{L}$ is

$$p(\mathbf{L}|\mathbf{Y}) \propto |\mathbf{L}^{-1}|^n p(\mathbf{X} = \mathbf{L}^{-1}\mathbf{Y}) p(\mathbf{L}) \tag{9}$$

To estimate the illuminant with minimum risk one simply places a grid over all admissible illuminants and computes the posterior mean in a single loop over all those illuminants in the grid.

2.1. The Greyworld algorithms

The greyworld algorithms as well as the more general class of grey-edge algorithms [16] are special instances of the formalism presented above. Greyworld algorithms and grey-edge algorithms are parameterized in the following form

$$\left(\sum_{i=1}^{N} \left| \frac{\partial^n f_i(x)}{\partial x^n} \right|^p \right)^{\frac{1}{p}} = kl. \tag{10}$$

The illuminant is estimated up to a scalar constant k which is independent of its color. In this approach three parameters need to be set, the derivative order n, usually in the range $n = 0, 1, 2$, its width σ, and the norm parameter p. All Greyworld algorithms of this form can be seen to be special instances of the Bayesian framework presented above. Each algorithm corresponds to a reflectance distribution whose maximum likelihood estimate yields exactly the same illuminant. For example the scale-by-max or white-patch algorithm

$$l_c = \max_i y_c(i) \tag{11}$$

corresponds to a constant $p(\mathbf{X})$. Greyworld algorithms with no derivative $n = 0$ correspond to a reflectance distribution which is independent for each pixel and channel $p(\mathbf{x}) = \prod_{c\in\{r,g,b\}} p(x_c)$ of the following form

$$p(x_c) = \frac{\alpha_c p_c}{\Gamma(1/p_c)} \exp\left(-\frac{(\alpha_c x_c)^p}{p_c k^{p_c}} \right). \tag{12}$$

Here we give the most general model using a scale parameter α_c and a parameter p_c for each channel independently. In the Greyworld algorithms the parameters are set to $\alpha_r = \alpha_g = \alpha_b = 1$ and $p_r = p_g = p_b$. The general form of Equation (12) expands the set of Greyworld algorithm but needs more parameters to be set. If one however has an accurate model for the reflectances or for example believes that there evidence for different scale parameters α_c it is possible to build an algorithm based on the more general form.

One potential drawback of the grey-edge method is the following. It computes $y_i - y_{i+1}$. If we were to operate in log space, it would give: $log(y_i/y_{i+1}) = log(x(i)/x(i+1))$. This means in log space the observation is independent of the illuminant. In practice, images are converted from raw camera data to an image format which involves a gamma correction, which is similar in form to a log operation.

2.2. A Prior for the Greyworld algorithms

By noting that the class of Greyworld algorithms are special cases of the Bayesian approach it is now possible to equip them with a prior for the illuminant $p(\mathbf{L})$. Using the reflectance distribution of the Greyworld algorithms introduced in Section 2.1 together with an illuminant one searches for a maximum of Equation 9. By introducing the prior distribution the maximum posterior is no longer given in closed form as it was the case with the Greyworld algorithms. However all necessary statistics to solve for the maximum l can be precomputed which makes the optimization problem very efficient. Thus solving with gradient descent incurs almost no extra computational time.

In this approach, it is necessary to balance the strength of the reflectance distribution versus the illuminant prior. The Greyworld reflectance distribution incorrectly assumes all pixels are independent and consequently the likelihood (7) vastly overcounts the amount of information in the image, overwhelming the prior. In order for the prior to have an effect, we must reduce the overcounting in the likelihood. One approach is to use bin clipping as in [13]. Another approach, complimentary to bin clipping, is to dampen the likelihood with an exponent η, representing the fraction of independent pixels in the image. This leads to the posterior:

$$p(\mathbf{L}|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{L})^\eta p(\mathbf{L}), \tag{13}$$

The value of η was determined with cross validation in the experiments from the set $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. Usually the value selected was $\eta = 0.001$. The prior was obtained by fitting a Gaussian distribution to the training split of the new Color Checker Database (results reported as (CC)). We also tried to use a different set of illuminants, namely the illuminants from the Greyball datasets (reported as (GB)). However the inclusion of either prior is not found to perform statistically significantly different to each other or the Greyworld algorithm alone.

3. Training with known illuminant color

Training data is needed to estimate the parameters of the reflectance prior $p(\mathbf{X})$ and the illumination prior $p(\mathbf{L})$. In [13] the training data consisted of news photographs in which the true illuminant and reflectances were completely unknown. The illuminant was estimated by the scale-by-max algorithm and the resulting reflectances were assumed to be correct, potentially biasing the results.

In this paper, we remove this bias by training only on images with known illuminant color. Specifically, each image has an object with known albedo placed in it. We assume the reflectance of the object to be the albedo times the cosine of the angle of incident light: $\mathbf{x} = \mathbf{a}\cos(\theta)$. This gives the equation $\mathbf{y} = \mathbf{L}\mathbf{a}\cos(\theta)$. With $\mathbf{y}$ and $\mathbf{a}$ known, we can solve for $\mathbf{L}$ up to a scale factor. Since the angle of incidence is unknown, we cannot recover the exact brightness of the light source from this equation.

Now the problem is reduced to estimating the brightness of the light source for each training image. We have experimented with two approaches. In the first approach, we estimate $\mathbf{L}$ as above and then scale by w where

$$w = \max_c \max_i \frac{y_c(i)}{\ell_c} \tag{14}$$

This chooses the smallest possible brightness consistent with the constraint that all reflectances in the image must be less than 1.

In the second approach, we iteratively re-estimate $p(\mathbf{X})$ and w to make the reflectance distribution as compact as possible. Starting from the estimate (14) for each training image, estimate the reflectances $\mathbf{X} = w^{-1}\mathbf{L}^{-1}\mathbf{Y}$, and then compute the overall training set frequencies m_k from these reflectances. Now revisit each training image and re-estimate the brightness w so that the resulting reflectances align with the frequencies m_k. The cost function we minimize is:

$$f(w) = \sum_k (m_k - m'_k(w))^2 \tag{15}$$

where $m'_k(w)$ is the frequency of reflectance k in the training image after division by the brightness w.

3.1. Illumination Prior

We use a different illuminant prior than [13]. In [13], the illuminant prior was chosen to be uniform over a subset of illuminants. Instead we use the empirical distribution of the training illuminants. That is, during test time we compute the likelihood (7) for all training illuminants and then take a likelihood-weighted average in chromaticity space. This has the advantage of biasing the algorithm toward frequently occurring illuminants.

3.2. Indoor/outdoor separation

The typical illuminants and reflectances in outdoor images are significantly different from indoor images. Rather than lump the statistics together into one compromise distribution, we can instead learn separate distributions for each type and switch between them at test time. For example, if we have an indoor/outdoor detector then we can apply only the reflectance/illumination prior appropriate for that image type.

4. Datasets for Color Constancy

To our knowledge there seem to be two different large-scale datasets for the task of color constancy. The first dataset [2] consists of 30 images of constructed scenes taken under 11 different illuminant sources. For each illuminant source the spectral distribution is known. Thus the dataset is of high quality but does not represent the full variation of typical scenes as outdoor scenes are missing. The second dataset is introduced by Ciurea and Funt [4] which consists of more than 11.000 frames from video, 6490 being outdoor scenes and 4856 indoor scenes. A grey ball is mounted onto the camera tripod and appears in each image in the lower right corner, to be used as a color reference. This dataset has however the shortcoming that the images have been subjected to correction and are available in low resolution only. Additionally the frames are highly correlated due to the continuous recording of video. Arguably around 600 uncorrelated images can be extracted from the full set.

Given these problems, we have collected a new dataset of 568 images, with a wide variety of indoor and outdoor shots also including a number of portraits. The images were taken with two high quality DSLR camera (Canon 5D and Canon 1D) with all settings in auto mode, and pictures were stored in RAW format. In each image a MacBeth color checker chart was placed as a reference. We took care to place the chart so that it is illuminated by what we perceived as the most dominant illuminant of the scene. Some example images are shown in Figure 2. The exact position of the chart was hand labeled. The last row of the chart consists of six patches of achromatic reflectance. Those patches were used to calculate the illuminant of the scene omitting very dark and too bright pixels.

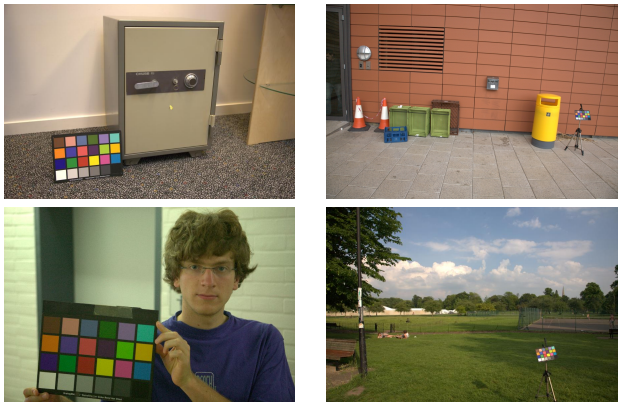

Figure 2. Example images from the Color Checker database. A MacBeth Color Checker chart is placed in the image so that it is illuminated by the most dominant illuminantion source.

The images are available in RAW format and all white-balance multipliers of the camera are stored alongside. Therefore it is possible to convert the images according to any of the different white balancing modes the cameras provide. For the experiments we used the Canon Digital Photo Professional program to convert the images into tiffs and rescaled them to 20% of their original size (813×541 and 874×583). We corrected for chromatic aberration and used the auto whitebalance setting of the Canon camera. We also checked with the program *dcraw* [5] and found that we obtain almost identical results.

5. Experiments

5.1. Error metrics for Color Constancy

There are several ways to measure the performance of color constancy algorithms, where ideally one is interested in a perceptual measure, unless one aims for color correction as a preprocessing step for feature extraction. Two metrics are used frequently, angular error and root mean squared error (RMSE). These two measures are independent of the brightness of the illuminant and simply compare colors. Angular error is the angle between the illuminant estimate $l = (l_r, l_g, l_b)$ and the ground truth illuminant vector for the image. RMSE is computed between the chromaticities of the illuminant $(l_r/(l_r + l_g + l_b), l_g/(l_r + l_g + l_b))$ and the ground truth illuminant.

For each algorithm we report three statistics of the error distribution. The mean RMSE of the predicted illuminant, the mean RMSE of the best 75% of the images ("Top75") and the mean RMSE of the worst 25% images ("Worst25"). For all algorithms we found the error distribution to be heavily skewed and thus use three measures to better represent its statistic. The "Worst25" measure is of particular interest for color correction applications, since it is the grossest errors which are least acceptable to the viewers of photographs.

5.2. Results on the fusion approach

Our first experiment with the new dataset re-examines the fusion approach proposed in [10]. The idea is as follows. From each training image a number of feature descriptors are computed. The set of all feature descriptors for all training images are subsequently clustered using K-means. For each cluster center one computes the performance of a set of proposal algorithms for all images whose descriptors fall into its Voronoi cell. The best performing algorithm is assigned to the cluster. During test phase one computes the features of the test image and searches for the closest cluster center. The illuminant of the test image is then estimated with the algorithm which is assigned to this cluster.

We follow exactly the protocol reported in [10]. Two derivative filters are applied to each color channel of each image, one in the horizontal and one in the vertical direction. A Weibull distribution is fitted to the statistics of these 6 images, and each distribution has two parameters, giving a 12-dimensional feature descriptor. We used K-means with 100 restarts and set the number of cluster centers K alternatively to be $5,\ 15$ or equal to the number of training images. We used the very same 5 algorithms as in[10] $((n, p, \sigma) = \{(0, 1, 0), (0, \infty, 0), (0, 13, 2), (1, 1, 6), (2, 1, 5)\}$ as proposal algorithms; however the conclusions should not depend too precisely on this choice. Each experiment is repeated 100 times and the RMSE and the we report the averaged RMSE and standard error.

The conclusions in [10] are based on the validation error from threefold cross validation, using the entire Greyball dataset. Since this dataset is highly correlated, the results reported in [10] are overly optimistic. The feature descriptors of consecutive frames of the video stream are likely to be very close and, given the random sampling schemes used for training and test, could lead to severe overfitting. That is indeed what happens, as can be seen in Table 1. The column "GB all" shows results using the whole dataset, and with sufficiently many cluster centers, the selection algorithm appears to beat the best fixed algorithm. (This is a single algorithm, not selected according to features of the data, but simply chosen as the algorithm that performs best on average over the training set.) However, to test the approach fairly, one has to restrict the dataset to a subset of independent images, so we take 500 images, evenly distributed throughout the videos. Using this "GB 500" dataset, the performance of the fusion approach drops severely (bottom of second column of the table), and now performs no better than the best fixed algorithm. To further investigate if the Weibull parameters are discriminative for the selection of different algorithms we plotted the parameters of the Weibull distribution of the first derivative image in Figure

	GB all	GB 500	CC
performance bound	23±0.2	23±0.2	41±1.5
best fixed algorithm	45±0.3	46±1.4	60±2.3
K = 5	43±0.3	44±1.5	57±2.0
K = 15	42±0.3	44±1.5	57±2.0
K = #training images	35±0.3	44±1.6	59±2.1

Table 1. Results of the fusion approach proposed in [10]. RMSE validation error with standard error (times 1000) is reported for threefold cross validation using the entire Greyball dataset (GB all), only 500 images taken at distant time steps (GB 500) and on the database introduced in this paper (CC). All numbers are averaged over 100 independent runs. The numbers in the "GB all" column are obtained by using the same experimental protocol as [10] and show clear evidence of overfitting — see text.

3. For each image we compute the best performing algorithm from the proposal which is coded with the markers in the Figure. The same markers correspond to the same algorithms. If the features are discriminative one would see clusters of the same algorithms. Obviously this is not the case.[2]

Now the fusion experiment is repeated with the new Color Checker database, in which images are shot independently, and illuminants are accurately labelled. The last column of Table 1 shows again that no benefit is gained from algorithm selection, since the improvement is not statistically significant.

In other words, no benefit at all is obtained from selecting an algorithm according to image features.

Note that a perfect selection method could, in principle, obtain a substantial performance gain (top row of Table 1, "performance bound"). However, no beneficial selection procedure has been obtained as yet. In seeking such a procedure, we tried features based on RGB histograms and rg chromaticity information to guide the selection but this did not improve the results using the K-means selection scheme (results not reported here). We also tried using the entire set of Weibull and other features to choose between the two best performing Greyworld algorithms with a discriminative approach using an SVM, but the results were never significantly better than the best fixed algorithm.

5.3. Results of Bayesian estimation

In this section we use the new Color Checker dataset to revisit the Bayesian estimation algorithm of Rosenberg et al. [13] to see whether its claimed superiority over Greyworld algorithms persists when accurate illumination labels and a large uncorrelated set of data are available.

[2]We also checked with more informative projections obtained by Linear Discriminant Analysis but observe the same behaviour. This plot is not to be confused with the ones presented in [10]. Those plots are color coded by first clustering the data and subsequently assigning an algorithm to each Voronoi cell.

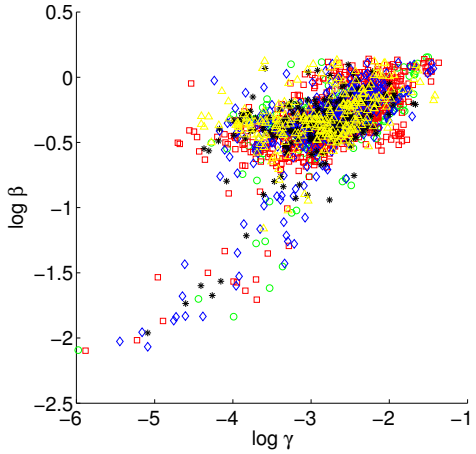

Figure 3. The two parameters of a Weibull distribution which are used in the fusion algorithm. Each different marker type corresponds to a Greyworld algorithm. If the Weibull parameters are informative one would see clusters of the same markers.

Before considering experiments with Greyworld algorithms, we first provided two baseline algorithms as sanity checks. The first is the built in color-correction algorithm for the Canon camera ("AutoWB"). The second is the trivial algorithm which simply returns a constant standard illuminant independent of the input image, this standard being computed as a mean over the training set illuminants. Two separate means are computed, one for indoor and one for outdoor images. We term this trivial algorithm "Mean Prediction". Since both also depend on the transformation used to convert the RAW images to RGB we tried both *dcraw* and the Canon converter shipped with the camera and found that both qualitatively give the same results.

Now Greyworld algorithms for comparison with Bayesian methods are chosen as follows. We created a proposal set of Greyworld algorithms by varying the parameters $n = 0, 1, 2$, $p = 1, 5, 20$ but fixed the derivative width for the GreyEdge algorithms to be 1. For outdoor images, we select the single Greyworld algorithm which best estimates illuminants for the training data, on average. For indoor images the algorithm selected is $n = 0, p = 1$ and for outdoor scenes the one with $n = 2, p = 5$. In addition we tested versions of the Greyworld algorithms with the addition of learned illuminant priors as in Section 2.2. For all experiments we perform three fold cross validation and report the validation error in Figure 4 (with further details in Table 2). All experiments are run separately on indoor and on outdoor test images.

For the Bayesian algorithm the training set is used to estimate the reflectance distribution as in Section 3. The ground truth set of training illuminants is used as a point-set representation of the prior distribution for illumination, separately over indoor and outdoor scenes. The results of these experiments are reported in Figure 4 as "Bayes (GT)". For comparison we applied the Bayesian algorithm using

the scale-by-max algorithm (SBM) proxy for illumination as in [13] and results are reported as "Bayes (SBM proxy)" in the figure. As a further experiment we obscured, in the training set, the indoor/outdoor image labels so that a single reflectance distribution was learned on both indoor and outdoor images jointly, and also used a joint indoor/outdoor illumination prior. Results are reported under the heading "Bayes (no indoor/outdoor)".

In addition, two variations on the Bayesian estimation were investigated that, it was thought, might improve the quality of estimates. First we replaced the clipping function (4) with a soft clipping function given in Equation (6), in order to improve the robustness of likelihood computation. We tried parameter values $\lambda = 0.001, 0.01, 0.1, 1, 2, 5$ and again report three fold cross validation results. For all folds on indoor images the best λ was found to be 0.001 which corresponds to almost no clipping. Outdoor images exhibit the opposite behaviour with best performance with high values of $\lambda = 1, 2, 5$. Results are presented "Bayes (tanh)". Next we experimented with re-estimating the brightness of the illuminant in order to sharpen the estimated reflectance distribution. This was done as described in Section 3 by minimizing Eq.(15). The results are reported as "Bayes (re-estimate)".

The following conclusions can be drawn from the experimental results in Figure 4, paying due attention to error bars and the statistical significance of claims.

1. In all cases the Bayes GT algorithm outperforms the baseline algorithms, the Canon auto white-balance (AutoWB) and trivial algorithm that returns mean illumination whatever the input image.

2. For outdoor images, Bayesian estimation with the new training data (Bayes GT) performs significantly better than Greyworld, even when Greyworld has the benefit of the illuminant prior.

3. Also on outdoor images, Bayes GT is significantly better than the Bayesian algorithm trained only with the proxy for illumination (SBM proxy).

4. On indoor images, the new Bayes GT algorithm with soft clipping performs significantly better than Bayes with only proxy illuminant labels, but does not significantly outperform Greyworld.

5. Two variations on the Bayes GT algorithm are tested — iterative re-estimation of illuminant to determine its magnitude and softened clipping in the evaluation of likelihoods — but neither are found to affect performance to a statistically significant degree.

6. Obscuring indoor/outdoor labels in the training set may have reduced performance (Bayes (no indoor/outdoor) relative to Bayes GT) but the effect is not statistically significant. However even with the obscured labels, the Bayesian estimator performs significantly better than the Greyworld algorithms on outdoor images, even when the Greyworld algorithm has the benefit of separate training on indoor and outdoor images.

6. Discussion and future work

The experiments reported have characterised thoroughly the performance of Bayesian estimation of illumination, relative to Greyworld algorithms. This study is particularly authoritative because, for the first time, image data is uncorrected and has accurate illumination labels, thanks to our new Color Checker dataset. We have confirmed that Bayesian estimation significantly outperforms Greyworld algorithms. We have demonstrated that accurate illumination labels significantly enhance the performance of Bayesian illumination estimation. In all cases, improvements over the manufacturer's built-in correction algorithm is substantial. In the future we hope to extend the Bayesian framework further by exploiting object detection.

Acknowledgments

We would like to thank the reviewers for many helpful comments. This work was mainly done during an internship of the first author at Microsoft Research, Cambridge. The first author was funded in part by the EC project CLASS, IST 027978.

References

[1] K. Barnard, L. Martin, A. Coath, and B. Funt. A comparison of computational color constancy algorithms - part 2. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(9):985–996, 2002.

[2] K. Barnard, L. Martin, B. Funt, and A. Coath. A dataset for color research. *Color research and application*, 27(3):147–151, 2002.

[3] D. Brainard and W. Freeman. Bayesian color constancy. *J. Optical Soc. of America A.*, 14(7):1393–1411, 1997.

[4] F. Ciurea and B. Funt. A large image database for color constancy research. In *Proc. Color Imaging Conference*. 2003.

[5] D. Coffin. Raw photo decoder "dcraw" v8.77 http://cybercom.net/ dcoffin/dcraw/.

[6] G. Finlayson, S. Hordley, and P. Hubel. Color by correlation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23:1209–1221, 2001.

[7] G. Finlayson and E. Trezzi. Shades of gray and colour constancy. In *Proc. IST Conf. Color Imaging*, pages 37–41, 2004.

[8] D. Forsyth. A novel algorithm for color constancy. *Int. J. Computer Vision*, 5:5–36, 1990.

[9] B. Funt and G. Finlayson. Color constant color indexing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(5):5–36, 1995.

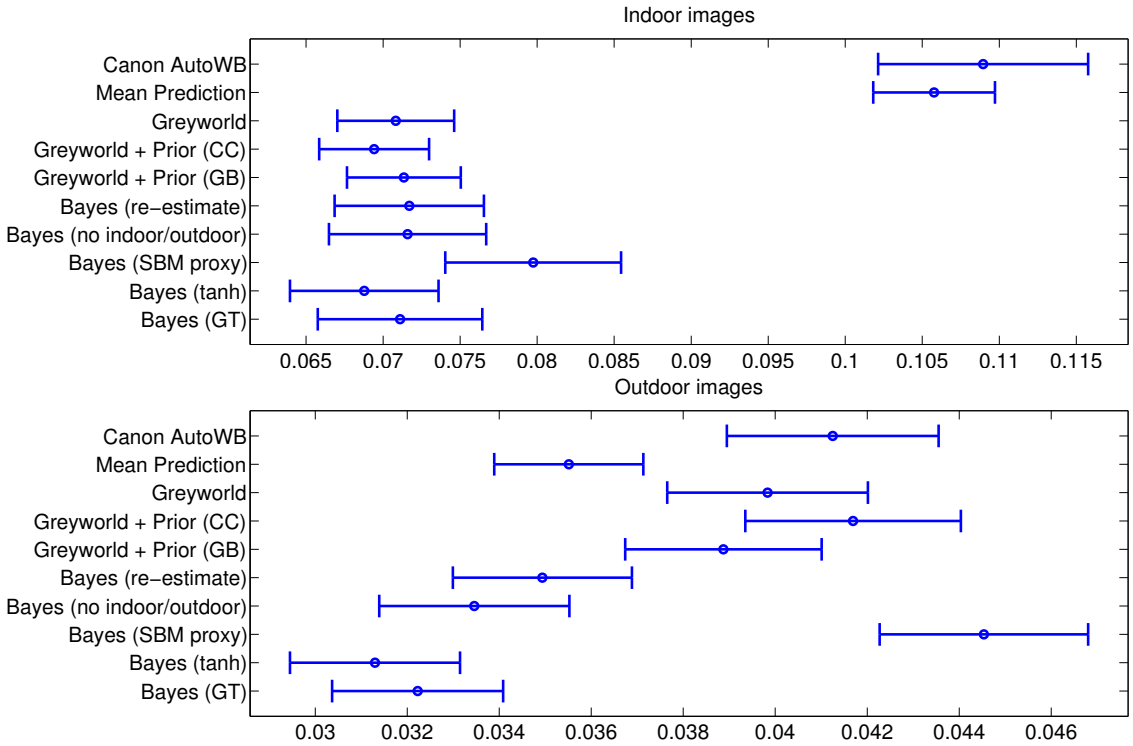

Figure 4. Results of different color constancy algorithms. The mean RMS error is plotted with scaled standard error. such that if two error bars do not overlap the difference is significant at the 95% level. All results are validation errors using three fold cross validation. Note that the x-axis of the upper and lower plot are scaled differently. Further details are given in Table 2.

	Indoor images			Outdoor images		
Method	RMSE	Top75	Worst25	RMSE	Top75	Worst25
Canon AutoWhiteBalance	109 ± 6	68 ± 4	235 ± 5	41 ± 2	25 ± 1	90 ± 3
Mean Prediction	106 ± 3	83 ± 2	174 ± 4	36 ± 1	25 ± 1	69 ± 3
Greyworld (performance bound)	48 ± 2	30 ± 2	103 ± 3	30 ± 2	16 ± 1	72 ± 3
Greyworld	71 ± 3	50 ± 2	135 ± 5	40 ± 2	24 ± 1	87 ± 3
Greyworld + CC prior	69 ± 3	49 ± 2	130 ± 5	42 ± 2	25 ± 1	92 ± 3
Greyworld + GB prior	71 ± 3	51 ± 2	135 ± 5	39 ± 2	24 ± 1	85 ± 3
Bayes (re-estimate)	72 ± 4	43 ± 2	158 ± 8	35 ± 2	21 ± 1	77 ± 3
Bayes (no indoor/outdoor)	72 ± 4	41 ± 2	165 ± 8	33 ± 2	19 ± 1	78 ± 3
Bayes (SBM proxy)	80 ± 5	44 ± 2	188 ± 8	45 ± 2	28 ± 1	93 ± 3
Bayes (tanh)	69 ± 4	39 ± 2	160 ± 7	31 ± 2	18 ± 1	71 ± 3
Bayes (GT)	71 ± 4	39 ± 2	167 ± 9	32 ± 2	19 ± 1	73 ± 2

Table 2. The numeric values of the results as presented in Figure 4. All results are cross validation error in RMSE times 1000 with standard error.

[10] A. Gijsenij and T. Gevers. Color constancy using natural image statistics. In *Proc. Conf. Computer Vision and Pattern Recognition*, Minneapolis, USA, June 2007.

[11] D. Jameson and L. Hurvich. Complexities of perceived brightness. *Science*, 133:174–179, 1961.

[12] E. Land and J. McCann. Lightness and retinex theory. *J. Optical Soc. of America A.*, 61:1–11, 1971.

[13] C. Rosenberg, T. Minka, and A. Ladsariya. Bayesian color constancy with non-Gaussian models. In *Advances in Neural Information Processing Systems*, Cambridge, MA, 2004. MIT Press.

[14] H. Trussell and M. Vrhel. Estimation of illumination for color correction. In *Proc. ICASSP*, 1991.

[15] J. van de Weijer, C. Schmid, and J. Verbeek. Using high-level visual information for color constancy. In *Proc. Int. Conf. on Computer Vision*, 2007.

[16] J. van de Weijer and T.Gevers. Color constancy based on the grey-edge hypothesis. In *Proc. IEEE Int. Conf. Image Processing*, volume 2, pages 722–725, 2005.