DEEP MULTI-RESOLUTION COLOR CONSTANCY

Caglar Aytekin[1], Jarno Nikkanen[2] and Moncef Gabbouj[1]

[1] Laboratory of Signal Processing, Tampere University of Technology, Tampere, Finland
[2]INTEL Finland, Tampere, Finland

ABSTRACT

In this paper, a computational color constancy method is proposed via estimating the illuminant chromaticity in a scene by pooling from many local estimates. To this end, first, for each image in a dataset, we form an image pyramid consisting of several scales of the original image. Next, local patches of certain size are extracted from each scale in this image pyramid. Then, a convolutional neural network is trained to estimate the illuminant chromaticity per-patch. Finally, two more consecutive trainings are conducted, where the estimation is made per-image via taking the mean (1st training) and median (2nd training) of local estimates. The proposed method is shown to outperform the state-of-the-art in a widely used color constancy dataset.

Index Terms— Color constancy, illuminant chromaticity estimation, deep learning, multi-resolution, local estimation.

1. INTRODUCTION

Color constancy is a unique feature of the human visual system, which enables robust perception of an object's true color under changing illumination. Computational color constancy (CC) aims to simulate this feature via computational models. The common approach is first to estimate the color of the illuminating source, and then to discount for it.

Color constancy methods can be categorized as unsupervised and supervised ones according to the way that they estimate the illuminant chromaticity. Several unsupervised methods rely on some assumptions on the scene reflectance statistics. Among these methods, White Patch (WP) [1] assumes that there is a perfectly reflecting object in the scene; Gray World (GW) [2] assumes that the average reflectance in a scene is gray; while Shades of Gray (SoG) [3] assumes that the average chromaticity in the scene is gray when raised to the power of p. This method experimentally searches for the optimal value of p. Gray Edge (GE) [4] algorithm on the other hand assumes that the average reflectance derivatives are gray. Such assumptions were shown to be derived from a single formulation with different parameters. Another group of unsupervised CC algorithms makes assumptions on the physical properties of objects. These methods focus particularly on specular objects in the scene, i.e. objects acting as mirror-like reflectors. In [6], a histogram based decomposition of reflected light into specular and illumination components was investigated. Lee [5] achieves this decomposition in CIE color space and using these results, the works in [7] and [8] show that specular components can be effectively exploited to estimate the color of the light source. Although providing a somewhat general approach to CC, the performance of unsupervised methods is somewhat limited due to the underlying assumptions.

Supervised approaches to color constancy either aim to learn a combination of unsupervised CC methods or learn the light source chromaticity directly. Combination-based methods exploit information such as semantic content of the scene [9] or scene type, e.g. indoor/outdoor scene [10]. Nevertheless, these methods are still dependent on the assumptions of unsupervised methods, as they only learn a combination of unsupervised methods.

Direct supervised methods are free from the assumptions of the unsupervised methods and learns a direct mapping from the input image to the illumination chromaticity. This is achieved by exploiting well-known machine learning algorithms such as neural networks [11], support vector regression [12] and Bayesian framework [13],[14]. Recently, due to their success in other fields, convolutional neural networks are also applied to CC and have shown promising results [15]-[16]. In [16], a three-stage learning is employed. First, a deep network was trained for object classification task. Second, this pre-trained network was fine-tuned to regress to an unsupervised CC method's output. Finally, the network parameters obtained by the second training were fine-tuned to regress to ground truth chromaticities. The main drawback of this study is the sub-optimal training procedure. The adaptation of the first deep network, dedicated to object classification, to CC task is suboptimal as the regression is made to the results of an unsupervised method. Moreover, the third network is trained only with thousands of images whereas the first network for classification task is trained with millions. In [15], a more effective use of deep networks is made via directly learning the CC problem through an 8-layer CNN that is trained on 32x32 patches. Considering the high-resolution datasets as in [17], this approach enables a significant population of the dataset from of the order of thousands to millions, making it more suitable for deep learning purposes.

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

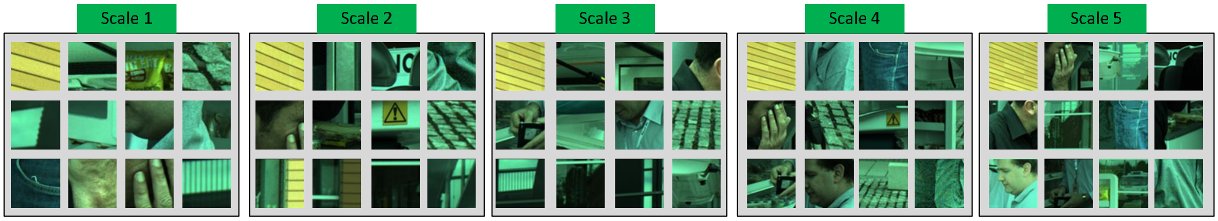

Fig. 1 Examples of extracted patches from each scale.

In [15], the network learned per-patch from this populated dataset is later fine-tuned per-image such that the knowledge of the global estimation from local patches is incorporated in the training. Despite the fine-tuning procedure, the initial training on 32x32 patches indirectly affects the final network's performance due to the possible limited scene contents in such small patches for accurate illumination estimation.

In this study, we propose a CNN-based approach to CC by addressing the shortcomings of the previous methods. First, we create an image-pyramid with progressively downscaling the original image at several scales. Then, local patches were extracted from all resolutions, which were then used as training samples. Such an approach eliminates the effect of limited content in small patches to some extent.

Moreover, the dataset is populated in greater numbers, which allows training a very deep network in an end-to-end manner. As expected, exploiting such a deep network helps to achieve better generalization accuracy.

The rest of the paper is organized as follows. In Section 2, the proposed method is introduced while the experimental results and discussion are provided in Section 3. Finally, Section 4 concludes the paper.

2. PROPOSED METHOD

2.1. Dataset Population

A limiting issue about application of deep learning to color constancy (CC) problem is the small size of available datasets. For example, a widely-used database, Gehler-Shi [17], consists of only 568 images. However, if one assumes that there is only one illuminating source in a scene, an image can be cropped to many local patches all of which have the same ground truth, i.e. the chromaticity of the single illuminating source. Therefore, the small number of high-resolution images can be populated to a much greater number of local patches that can be exploited in deep learning based approaches [15]. The problem with this approach is the fact that local patches may sometimes correspond to regions that only cover very limited contextual information. Therefore, one should be aware of the tradeoff between data population and loss of useful context information when selecting the size of local patches. We propose handling this tradeoff as follows. We keep the original size of the images in Gehler-Shi dataset and form an image pyramid of 5 images by progressively scaling the original image 5 times with 0.75 scale. From each image in the pyramid, we extract 64x64 patches and form a database out of these patches. This way, we make use of a wide variety of scales and especially in the higher scales, the 64x64 local patches are not expected to correspond to regions with limited context as they actually corresponding to much larger regions in the original image. This can be observed from Fig. 1 where the local patch samples from each scale is illustrated. As expected, the patches in higher scales cover a large variety of context, hence possess a richer color distribution. The above dataset population results in 833007 patches which is suitable to train very deep architectures.

2.2. Preprocessing

The images in the Gehler-Shi dataset are taken by two cameras, one with black level 0 and another with 129. This black level is extracted from the images taken by the second camera. As it is a common procedure in deep learning, the 12 bit images are normalized such that the maximum possible intensity is 1. A global histogram stretching is also applied to each image in order to gain robustness to illumination intensity. The stretching is only applied to the illumination channel in HSV color space. Note that the above preprocessing steps do not distort the chromaticity of the pixels, hence do not degrade the performance of illuminant chromaticity estimation.

2.3. Network Architecture

The convolutional neural network exploited consists of three blocks of convolution, activation and pooling layers. The convolution filters are selected to be of size 5x5, 5x5 and 4x4 in the first, second and third layers, respectively. In all layers, the number of convolutional filters are selected as 32. The activation functions are rectified linear units, which are commonly used in deep learning. All pooling operations are performed by selecting the maximum in a 2x2 patch with stride of two. Finally, a fully connected layer is employed in the end with 256 hidden neurons. An illustration of the network is provided in Fig. 2. The proposed network consists of 6 weighted layers and has approximately 160000 parameters.

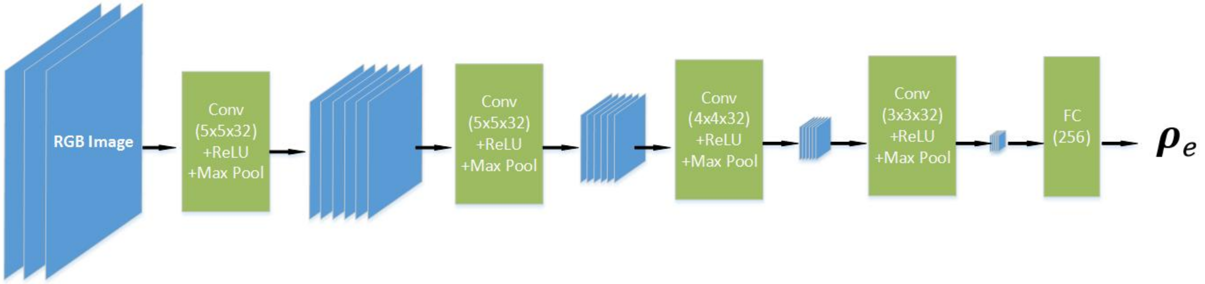
RGB Image
Conv (5x5x32) +ReLU +Max Pool
Conv (5x5x32) +ReLU +Max Pool
Conv (4x4x32) +ReLU +Max Pool
Conv (3x3x32) +ReLU +Max Pool
FC (256)
$\boldsymbol{\rho}_e$

Fig. 2 The proposed convolutional neural network architecture.

Fig. 3 Representative Examples from the Gehler-Shi Dataset **[17]**.

2.4. Loss and Error Functions

Training is conducted via backpropagating the Euclidean difference (the loss function) between the ground truth illumination chromaticity $\boldsymbol{\rho}_{GT}$ and the estimated one by CNN, $\boldsymbol{\rho}_e$. The error function on the other hand is the commonly used recovery angular error (RAE) calculated as in Eq. (1). The Euclidean loss is a good approximation of the RAE. In fact, given that $\boldsymbol{\rho}_{GT}$ and $\boldsymbol{\rho}_e$ are l_2 normalized, minimizing the Euclidean loss exactly corresponds to minimizing RAE.

$$RAE(\rho_{GT}, \rho_e) = \cos^{-1}\left(\frac{\boldsymbol{\rho}_{GT} \cdot \boldsymbol{\rho}_e}{\|\boldsymbol{\rho}_{GT}\|\|\boldsymbol{\rho}_e\|}\right) \tag{1}$$

2.5. Three-stage Training

In training, 3-fold cross validation is utilized based on the folds suggested by the dataset. We employ a three-stage training strategy. The first training is applied directly on the 64x64 image patches with batch size 256. Due to the large number of patches available, this training helps achieving effective training of CNN.

The second training is applied on the 568 samples (original images) where the estimated illumination chromaticity is determined by taking the prediction averages of all 64x64 patches corresponding to the image at hand. This training is utilized via fine-tuning the parameters of the CNN obtained by the first training.

The third training is similar to the second with the only difference that the median of the predictions from the local patches is taken as the final estimate. The third training fine-tunes the CNN parameters obtained by the second training.

During testing, the model obtained by the third training is used to estimate the illumination of each 64x64 patch corresponding to an image and the final estimate is the median of these local estimates.

3. EXPERIMENTAL RESULTS

The experiments have been conducted on Gehler-Shi dataset [17]. A set of representative images from this dataset is provided Fig. 3. As can be seen, the images cover a wide variety of real-world scenarios including images from indoor/outdoor scenes as well as images with limited/rich color distribution. It is worthy to note here that the images in Fig. 3 are color corrected with the ground-truth and are only shared for visualizing the contents in the dataset. The color-checker presented in these images are masked during the training and any 64x64 patch covering this masked area have simply been discarded.

Next, we evaluate the proposed method and compare it against several statistics- and learning-based methods. In the comparisons, we use the mean, median and maximum of the error metric in Eq. (**1**) across the dataset.

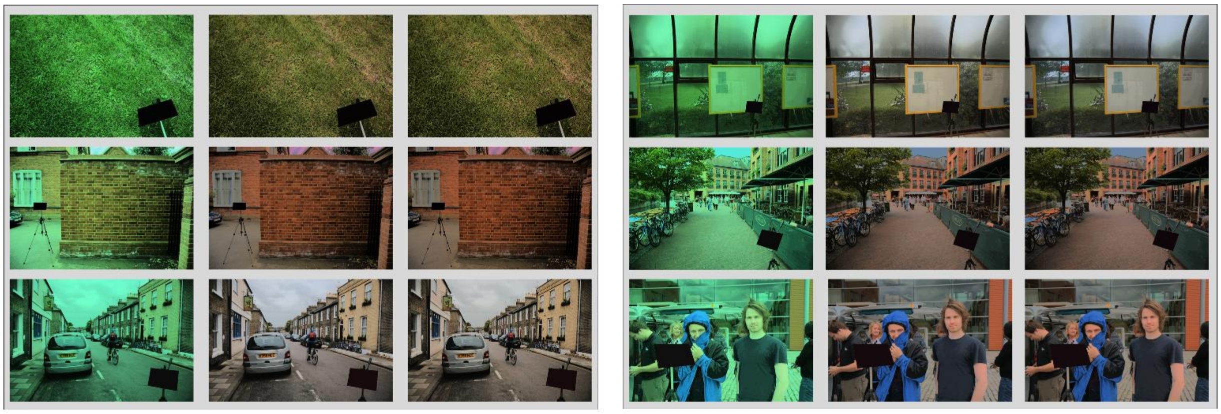

Fig. 4 (Left) Original image, color corrected images (middle) with our method, (right) with respect to ground truth.

Table 1 Comparison of the methods based on median, mean and maximum recovery angle error.

	Median	Mean	Maximum
Doing Nothing	13.55	13.63	27.37
White Point [1]	5.61	6.27	40.59
Gray World [2]	6.27	6.27	24.84
Shades of Gray [3]	4.04	4.85	19.93
First Order Gray Edge [4]	4.55	5.21	19.69
Second Order Gray Edge [4]	4.43	5.01	16.87
Gamut Mapping (pixel) [18]	2.30	4.20	23.20
Gamut Mapping (edge) [18]	5.00	6.50	29.00
Bayesian [14]	3.44	4.70	24.47
Natural Image Statistics [19]	3.13	4.09	26.20
Spatial-spectral (GP) [21]	2.90	3.47	14.80
Spatial-spectral (ML) [21]	2.93	3.55	15.25
Exemplar [20]	2.30	2.90	19.40
SVR [12]	6.67	7.99	26.08
CNN [16]	2.30	3.10	-
CNN [15]	1.98	2.63	14.77
Proposed-Training 1	1.83	2.61	14.02
Proposed-Training 2	1.57	2.47	13.77
Proposed-Training Final	**1.51**	**2.46**	**13.74**

It can be observed from Table 1 that the proposed method achieves the top performance in all evaluations. Moreover, the effect of the three-stage training is evident from the progressive performance increase in the test set after each training.

In Fig. 4, we also share a visual evaluation where we provide some examples of original images and color corrected images with respect to the estimated illuminant chromaticity with our proposed method and with respect to the ground truth. All images in Fig. 4 are obtained via global histogram stretching described in Section 2.2 for illustration purposes. It is also evident from Fig. 4 that the proposed method achieves a good estimation of the true illuminant chromaticity in a wide range of images containing indoor and outdoor scenes with very visually complex scenes.

Since the proposed method uses all possible patches from all resolutions, one might expect the computational complexity to be heavy. Since the method is highly parallelizable, this enables getting the final estimate in a reasonable time. In an un-optimized Matlab implementation with GPU (NVIDIA Quadro K2000M) processing, for a typical image of resolution 2193x1460, the proposed method takes 5.93 seconds on average to get the final estimate.

4. CONCLUSION

In this paper, an application of deep learning in a low-level image processing task, namely color constancy was investigated. From the discussion in the text, one can conclude that the proposed multiresolution approach is helpful in two ways: (1) local patches extracted from higher scales are more likely to include a rich color distribution which helps illuminant estimation, (2) it increases the number of available patches extracted from an image, which in turn enables training of deeper networks. These conclusions are also supported by the state-of-the-art performance of the proposed framework. As a future work, we plan to use a weighted pooling of local estimates with respect to the resolution they belong. Such an approach would learn the contribution of each resolution to the final estimate, which would be useful to get a more accurate estimate.

11. REFERENCES

[1] E. H. Land and J. J. McCann, “Lightness and Retinex Theory,” JOSA, vol. 61, no. 1, pp. 1-11, 1971.

[2] G. Buchsbaum, “A spatial Processor Model for Object Color Perception,” vol. 310, no. 1, pp. 1-26, 1980.

[3] G. D. Finlayson and E. Trezzi, “Shades of Gray and Colour Constancy,” in Proc. CIC, 2004, pp. 37-41.

[4] J. Van de Weijer, T. Gevers, A. Gijsenij, “Edge-based Color Constancy,” IEEE Trans. Image Process., vol. 16, no. 9, pp 2207-2214, 2007.

[5] H. C. Lee, “Method for Computing the Scene-Illuminant Chromaticity from Specular Highlights,” J. Opt. Soc. Am. A, vol. 3, pp. 1694-1699, 1986.

[6] G. J. Klinker, S. A. Shafer and T. Kanade, “The Measurement of Highlights in Color Images,” Int. J. Comput. Vision, vol. 2, pp. 7-32, 1988.

[7] G. D. Finlayson and G. Schaefer, “Solving for Color Constancy using a Constrained Dichromatic Reflection Model,” Int. J. Comput. Vision, vol. 42, no. 3, pp. 127-144, 2001.

[8] R. T. Tan and K. Ikeuchi, “Separating Reflection Components of Textured Surfaces Using a Single Image,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 2, pp. 178- 193, 2005.

[9] J. Van de Weijer, C. Schmid and J. Verbeek, “Using High-level Visual Information for Color Constancy,” in Proc. IEEE ICCV, 2007, pp. 1-8.

[10] S. Bianco, G. Ciocca, C. Cusano and R. Schettini, “Improving Color Constancy Using Indoor-Outdoor Image Classification,” IEEE Trans. Image Process., vol. 17, no. 12, pp. 2381-2392, 2008.

[11] V. C. Cardei, B. Funt and K. Barnard, “Estimating the Scene Illumination Chromaticity by Using a Neural Network,” J. Opt. Soc. Am., vol. 19, no. 12, 2002.

[12] B. Funt, W. Xiong, “Estimating Illumination Chromaticity via Support Vector Regression,” in Proc. CIC, 2004, pp. 47-52.

[13] C. Rosenberg, A. Ladsariya and T. Minka, “Bayesian Color Constancy with non-Gaussian Models,” in Proc. NIPS, 2003.

[14] P. V. Gehler, C. Rother, A. Blake, T. Minka and T. Sharp, “Bayesian Color Constancy Revisited,” in Proc. CVPR, 2008, pp. 1-8.

[15] S. Bianco, C. Cusano, R. Schettini, “Color Constancy Using CNNs,” in Proc. IEEE CVPRW, 2015, pp. 81-89.

[16] Z. Lou, T. Gevers, N. Hu and M. Lucassen, “Color Constancy by Deep Learning,” in Proc. BMVC, 2015.

[17] L. Shi and B. Funt, "Re-processed Version of the Gehler Color Constancy Dataset of 568 Images," accessed from http://www.cs.sfu.ca/~colour/data/

[18] K. Barnard, “Improvements to Gamut Mapping Colour Constancy Algorithms,” in Proc. ECCV, 2000.

[19] A. Gijsenij and T. Gevers, “Color Constancy Using Natural Image Statistics and Scene Semantics,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 4, pp. 687-698, 2011.

[20] H. R. V. Joze, M. S. Drew, “Exemplar-based Color Constancy and Multiple Illumination,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 5, pp. 860-873, 2014.

[21] A. Chakrabarti, K. Hirakawa, and T. Zickler, “Color constancy with spatio-spectral statistics,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 8, pp. 1509-1519, 2012.