https://doi.org/10.2352/ISSN.2470-1173.2017.14.HVEI-133

Evaluation of Color Prediction Methods in Terms of Least-Dissimilar Asymmetric Matching

Emitis Roshan and Brian Funt, School of Computing Science, Simon Fraser University, Vancouver, British Columbia, Canada

Abstract

The performance of color prediction methods CAT02, KSM^2, Waypoint, Best Linear, MMV center, and relit color signal are compared in terms of how well they explain Logvinenko & Tokunaga's [1] asymmetric color matching results. In their experiment, given a Munsell paper under a test illuminant, 4 observers were asked to determine (3 repeats) which of 22 other Munsell papers made the least-dissimilar match under a match illuminant. Given this data, we address the following four questions. Question 1: Are observers choosing the original Munsell paper under the match illuminant? If they are, then the average (12 matches) color signal (cone response triple or XYZ) made under a given illuminant condition should correspond to that of the Munsell paper's color signal under the match illuminant. Computation shows that in 274 of the 400 cases, the relit color signal is close to the mean color signal of the matches. Question 2: How do algorithm predictions compare to the average observer prediction of the actual color signal of the relit paper? The Wilcoxon signed-rank test shows that KSM^2, Waypoint, and Best Linear perform equally, and that both slightly outperform the observer average, which, in turn, significantly outperforms CAT02, and MMV (metamer mismatch volume) center. Question 3: Which method most closely predicts the observer average? We found that the color signal of the relit reflectance is a better predictor of the average observer than Best Linear, which in turn is marginally better than Wpt and KSM^2, both of which outperform CAT02 and MMV center. Question 4: Do the observers agree with one another? Using a leave-one-observer-out comparison shows that individual observers predict the average matches of the remaining observers somewhat better than the relit color signal, which in turn slightly outperforms Best Linear, Wpt and KSM^2, which then all significantly outperform CAT02 and MMV center.

Introduction

Logvinenko & Tokunaga [1] conducted an asymmetric color matching experiment in which observers view a Munsell paper under one light (the test illuminant) and then choose the least dissimilar matching paper from a set of 22 papers under a second light (the match illuminant). There were 4 observers and 3 repetitions each. The papers under both lights are always visible simultaneously. See Figure 1 for a photograph of the setup. The papers are rearranged between trials. Note that these are real papers under real illuminants, not colored patches on a digital display nor colors obtained using hidden illuminants to simulate reflectance changes [2] [3]. The experiment involved 6 illuminants of approximately equal illuminance and all 30 possible pairs were used as test/match illuminant conditions. However, since the two red illuminants are very similar, in this paper we exclude one of them and consider only the illumination conditions based on the 20 possible non-identical pairs of 5 of the illuminants.

The Logvinenko & Tokunaga (L&T henceforth) experiment differs from many other asymmetric color matching experiments in that subjects are not asked to match colors, but rather to choose the color which is least-dissimilar.

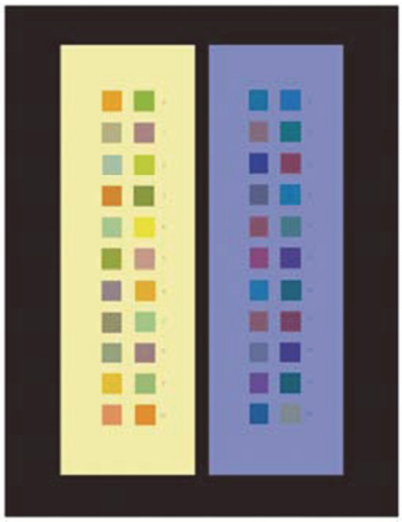

Figure 1. The asymmetric matching setup used by Logvinenko and Tokunaga [1] showing the example of the left-hand panel in yellowish light and the right-hand panel in bluish light. The papers are rearranged between trials.

During each trial, a laser pointer is used to select a test colored patch (a Munsell paper from the matte collection) from the left-hand panel and observers are asked to identify the least-dissimilar patch from the right-hand panel. As L&T point out, a perfect asymmetric match will usually be impossible due to metamer mismatching (i.e., the fact that two different reflectances may reflect metameric lights under one illuminant, but non-metameric lights under a second illuminant). Further analysis of the effect of metamer mismatching in the context of this experiment is provided by Logvinenko et al. [4].

Based on the L&T asymmetric matching results, we compare several color prediction methods to determine which best models observer performance. In particular, we compare von-Kries-rule-based CAT02 [5], KSM^2 [6], Wpt [7], Best Linear 3x3 transform [8], MMV (metamer mismatch volume) center [9] and Relit color signal (LMS cone response or XYZ) of the test paper under the match illuminant. Details of these methods are given below. In all cases, we assume that the methods have accurate information about the test and match illuminants. These methods are divided into two groups: those that require the full spectral power distribution of the illuminants (Wpt, Best Linear, MMV center, Relit color signal), and those that require only the color signals of the illuminants (CAT02, KSM2).

In analyzing the methods relative to the L&T data, we address four questions: (i) Are observers generally choosing the original Munsell paper under the match illuminant? (ii) If the average color signal of the observers' least-dissimilar matches is considered as a prediction of the actual color signal of the relit paper then is it better or worse than the predictions made by the computational methods? (iii) Which computational method most closely corresponds to the observer average? and (iv) How does the

performance of individual observers compare to the computational methods in predicting the least-dissimilar matches of the average observer?

Background

Numerous methods for predicting 'color' under a change of illumination have been proposed. Derhak and Berns [7] make the distinction between chromatic adaptation transforms (CATs) and material adjustment transforms (MATs). A CAT is intended to predict what color signal under the match condition will appear the same as under the test condition. Of course there is the issue of what 'the same' means. Derhak and Berns define the goal of a MAT as "…to predict material constancy or how sensor excitations for an object color change with changes in observing conditions" [7]. The problem with this definition is, as established by Logvinenko et al. [4], that as a result of metamer mismatch intrinsic object colors that are independent of the illuminant simply do not exist—hence material constancy does not exist either. However, so long as we bear in mind that we will not obtain constancy or "material color equivalency" [7] we can still investigate methods of predicting—given a color signal from a given surface reflectance under a first light—what its color signal is likely to be under a second light. Wpt [7] is one such color signal predictor. However, the issue we address here is not whether one CAT or color signal predictor is better than another, but rather whether or not any of them successfully predict the least-dissimilar matches made by the observers in L&T's experiment.

Color signal predictors can be divided into two categories: those that require full knowledge of the spectral power distributions of both the test and match illuminants; and those that require only the color signals of the perfect reflector under each illuminant. In the first category are Relit, Best Linear [8], Wpt [7] and MMV center [9].

The Relit color signal is simply the color signal of the given test paper under the match (second) illuminant. Computing it requires the full spectral reflectance function of the surface as well as the SPDs of the second illuminant. Since L&T used matte Munsell papers, we assume that the color signal $(\varphi_1, \varphi_2, \varphi_3)$ resulting from light impinging on sensors $R_k(\lambda)$ (k = 1…3) from a surface of spectral reflectance S(λ) illuminated by light with spectral power distribution E(λ) is:

$$\phi_k(x) = \int_{\lambda_{min}}^{\lambda_{max}} S(\lambda)E(\lambda)R_k(\lambda)d\lambda \ (k = 1,2,3) \quad (1)$$

The Relit 'prediction' of the color signal is not a prediction but rather, under the assumption of matte reflectance, it is the actual answer. Wpt involves a 3x3 linear matrix transformation of the test color signal to the match color signal. The 3x3 transformation is determined based on the SPDs of the illuminants and a training set consisting of the reflectances of all the papers in the Munsell collection. In order to satisfy other design requirements, Wpt does not, in fact, determine the optimal 3x3 matrix. In comparison, the Best Linear method [8] is based on using the optimal 3x3 matrix mapping the color signals from the training set (1600 Munsell papers) under the test illuminant to the match illuminant.

MMV center prediction is based on computing metamer mismatch volumes. For a given color signal under the test illuminant, the set of color signals it could theoretically become under the match illuminant defines a convex volume in color signal space called the metamer mismatch volume (MMV). Computing the MMV requires full knowledge of the SPDs of both illuminants. Logvinenko et al. [9] propose using the color signal at the geometric center of the MMV as the prediction of what the color signal under the test illuminant is likely to become under the match illuminant.

In the second category of color signal prediction methods—those that require only the color signals of the illuminants—we consider von-Kries-based CIECAM02 [5] and KSM^2 [6]. At the heart of CIECAM02 is the chromatic adaptation transform CAT02, which applies the standard von Kries (diagonal) transformation after a sharpening transformation [10][11]. The sharpening transform is tuned on corresponding color datasets and therefore it is not specifically designed to predict color signals of surfaces under the test illuminant.

Also in the second category is KSM^2 developed by Mirzaei et al. [6]. KSM^2 uses Gaussian-like reflectance functions (called wraparound Gaussians). Each such Gaussian reflectance is specified by 3 parameters: K the scaling, S the sigma, M the peak wavelength. To make a color signal prediction, KSM^2 finds three Gaussian functions, one representing an SPD metameric to the test illuminant, a second metameric to the match illuminant, and a third representing a reflectance metameric to the given test color signal under the Gaussian SPD metameric to the test illuminant. It then computes the color signal of that Gaussian reflectance under the match Gaussian illuminant and uses that color signal as its prediction.

Observers choose original Munsell paper

Are observers generally choosing the original Munsell paper under the match illuminant as the least-dissimilar one? For each illumination condition, 4 observers with 3 repeats made least-dissimilar matches. All 20 chromatic papers were used as test papers. For each of the 20 test papers, therefore, there are 12 least-dissimilar matches reported. If the observers are selecting the original paper as being least-dissimilar then the color signal of the original paper under the match illuminant (the relit color signal) can be expected to be close to the average relit color signal of the selected papers. Figure 2 shows examples of the 95% confidence region centered on the average relit color signal of the observer matches for four of the test papers for the red-neutral illumination condition (i.e., red test, neutral match illuminant). In the plot, three of the four relit color signals fall within the 95% confidence interval.

Figure 2 shows the general trend of the confidence ellipses (or ellipsoids in XYZ) with the relit color signal falling in or near it. To check all 400 cases (20 papers under 20 illuminant conditions), we computed the Mahalanobis distance for each relit color signal from the mean observer color signal for each particular illumination condition. The advantage of the Mahalanobis distance measure is that it is independent of any full-rank linear transformation of the sensor space such as the Hunt-Pointer-Estevez transformation from XYZ to LMS. As a reminder, the Mahalanobis distance of a point $\vec{x} = (x_1, x_2, \ldots, x_n)^T$ to a set of points $\vec{y_i}$ with mean $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^T$ and covariance matrix S in n-dimensional space is defined as:

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})} \quad (2)$$

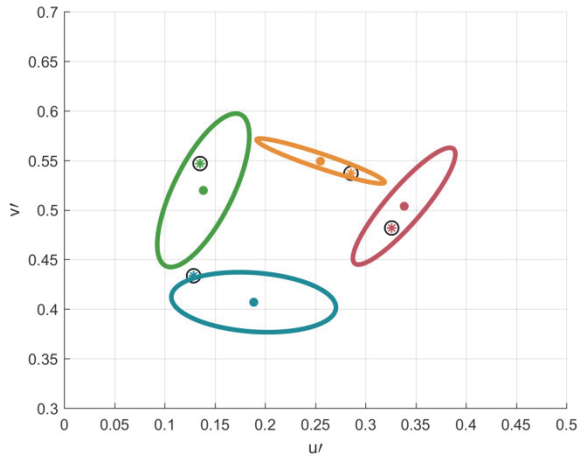

FIGURE 2. Four examples of the 95% confidence ellipses centered on the mean (dot) of the 12 matches with the locations of the relit color signals of the Munsell papers marked by asterisks surrounded by a ring. The illumination condition is red to neutral. Starting at the bottom ellipse and proceeding clockwise the ellipses correspond to Munsell papers 10BG 5/10, 5GY 7/12, 10R 5/16 and 5RP 5/12.

Since XYZ space is 3-dimensional, the expected distribution of Mahalanobis distances is a chi-squared distribution with 3 degrees of freedom. The critical values of the associated chi-squared distribution are used to check if a point at a given distance D_M is inside the given ellipsoidal confidence interval or not. However, since there are only 12 observations per patch/illuminant condition there are quite a few cases where all 12 observations fall within a 2-dimensional or, occasionally, a 1-dimensional subspace. This happens whenever all the observers agree on a set of only two or three candidate papers. In these cases, so long as the point is close to the subspace (as measured in terms of eigenvalues of the singular value decomposition) we project the point into the subspace before computing the Mahalanobis distances. It should be noted that if the points all fall on a plane then the expected distribution of Mahalanobis distances will be from a chi-squared distribution with 2 degrees of freedom and the confidence interval will be an ellipse instead of an ellipsoid based on the critical values for 2 degrees of freedom. Similarly, if the points fall on a line then the confidence region will turn into a simple confidence interval.

In 274 cases of the 400 cases (68.5%), the relit paper's color signal falls inside the corresponding confidence region. We also did the same test but based on CIE1976 u′v′ chromaticity coordinates and found that in 291 cases the relit color signal was within the corresponding confidence ellipse. These results suggest, perhaps not surprisingly, that observers generally (but definitely not always) find the match paper that is physically identical to the test paper to be the least dissimilar one.

Color Signal Prediction Comparison

Whether or not observers are actually predicting what the color signal of a given paper under the test condition will be under the match condition, we can, nonetheless, treat the average color signal of the least-dissimilar matches as a predictor and evaluate how accurate that predictor is. In the L&T experiment there were 4 observers so we have 5 different predictors. ObsC is the predictor based on combining all 12 results (4 observers, 3 repeats) by averaging them and Obs1, Obs2, Obs3, Obs4 are predictors based on the average of each observer's 3 repeats taken separately. For comparison, we also used implementations of KSM^2, MMV center, Wpt, CAT02 and Best Linear as predictors.

Table 1. Comparing Relit Color Signal Predictions. Column 3 indicates in how many cases out of the 20 different illumination conditions that the Wilcoxon test indicates that method from column 1 has a statistically lower error than the method from column 2 in predicting the color signal of the 20 papers under the match illuminant; column 4 the reverse; and column 5 when they are statistically equivalent.

Method 1	Method 2	Err1<Err2	Err2<Err1	Equal
ObsC	KSM^2	2	5	13
Obs1		6	3	11
Obs2		2	4	14
Obs3		0	7	13
Obs4		1	3	16
ObsC	CAT02	14	2	4
Obs1		14	0	6
Obs2		14	2	4
Obs3		9	2	9
Obs4		12	2	6
ObsC	MMV Center	19	0	1
Obs1		20	0	0
Obs2		20	0	0
Obs3		18	0	2
Obs4		19	0	1
ObsC	Wpt	0	6	14
Obs1		4	3	13
Obs2		2	5	13
Obs3		0	7	13
Obs4		0	5	15
ObsC	Best Linear	0	7	13
Obs1		0	5	15
Obs2		2	6	12
Obs3		0	9	11
Obs4		0	6	14
KSM^2	CAT02	18	0	2
KSM^2	MMV Center	20	0	0
KSM^2	Wpt	1	2	17
KSM^2	Best Linear	0	8	12
CAT02	MMV Center	14	0	6
CAT02	Wpt	0	20	0
CAT02	Best Linear	0	20	0
MMV Center	Wpt	0	20	0
MMV Center	Best Linear	0	20	0
Wpt	Best Linear	0	2	18

Using the same test/match conditions and Munsell papers as above, we compute the average prediction error for each of the predictors measured in terms of the Euclidean distance between the predicted XYZ and the relit XYZ. Although the results reported here are in terms of XYZ, almost identical ranking results were obtained using Euclidean distances in Hunter-Pointer-Estevez LMS space. We compare the performance of the predictors to one another using the Wilcoxon one-sided and two-sided tests. The Wilcoxon test results are tabulated in Table 1. All the tests are performed at the 5% significance level. Note that the three

rightmost columns of the table show the number of illumination conditions for which Method 1 is statistically better on average than Method 2, whether Method 2 is better than Method 1, or whether they are statistically equal.

Overall the results in Table 1 indicate Best Linear, Wpt and KSM^2 are roughly equivalent and all are slightly better predictors than ObsC; however, all four of them significantly outperform CAT02 and MMV center. Note that as mentioned above that Best Linear, Wpt, and MMV center require the full spectra of the test and match illuminants, while ObsC, KSM^2 and CAT02 require only their color signals. In other words, the former ones may or may not predict human performance, but they cannot possibly provide a computational model of any aspect of color perception.

Predicting Observer Average Matches

To determine which method most closely predicts observer least-dissimilar matching behavior, for each of the 20 illumination conditions, we use the 12 (4 observers, 3 repeats) matches made and compute the average of the color signals arising from the matched papers under the match illuminant. We then compare this average observer least-dissimilar match to the predictions made by the various computational methods measured in terms of Euclidean distance between the respective color signals. The Wilcoxon signed-rank test was then used to evaluate the methods with respect to one another in terms of their performance in predicting the 20 matches. The Wilcoxon results for the 20 illumination conditions are listed in in Table 2. As an example, the first row in the table indicates that KSM^2 outperforms the Relit color signal in only one illumination condition, the Relit outperforms KSM^2 in 4 illumination conditions, and in 15 illumination conditions their performance is evaluated as statistically equivalent.

Table 2. Comparison to Average Observer Matching. Column 3 indicates in how many cases out of the 20 different illumination conditions that the Wilcoxon test indicates that method from column 1 has a statistically lower error than the method from column 2 in predicting the average observer least-dissimilar matches of the 20 papers; column 4 the reverse; and column 5 when they are statistically equivalent.

Method 1	Method 2	Err1<Err2	Err2<Err1	Equality test
Relit	KSM^2	4	1	15
Relit	MMV Center	20	0	0
Relit	CAT02	9	0	11
Relit	Wpt	2	0	18
Relit	Best Linear	0	0	20
KSM^2	MMV Center	19	0	1
KSM^2	CAT02	5	2	13
KSM^2	Wpt	1	0	19
KSM^2	Best Linear	0	2	18
MMV Center	CAT02	0	17	3
MMV Center	Wpt	0	19	1
MMV Center	Best Linear	0	20	0
CAT02	Wpt	0	6	14
CAT02	Best Linear	0	9	11
Best linear	Wpt	0	0	20

In sum, the results in Table 2 indicate that the color signal of the relit reflectance and best linear fit estimator are equivalent predictors of the average observer, and both are only marginally better than Wpt and KSM^2, both of which, in turn, clearly outperform CAT02 and MMV center. KSM^2 is, in comparison to the other methods, both a good predictor and requires only the color signals of the illuminants, not their full spectral power distributions.

Note that the results in this Table 2 show the *relative* performance of the methods, not the absolute performance. In other words, the methods might be doing equally poorly rather than equally well. Table 3 lists the accuracy of each method's predictions averaged over the 400 cases. The accuracy is measured in terms of the Euclidean distance between the prediction and the average XYZ of the 12 least-dissimilar matches, and similarly for u′v′coordinates.

Table 3. Accuracy in Predicting Average Observer Matches Mean and median over the 400 cases of the Euclidean distance in XYZ and CIE1976 u′v′ between each method's predictions and the average observer match.

Method	Mean XYZ	Median XYZ	Mean u'v'	Median u'v'
Relit	5.21	3.45	0.024	0.015
Best Linear	5.56	4.17	0.040	0.023
Wpt	6.20	4.44	0.096	0.025
KSM^2	8.08	4.50	0.043	0.030
CAT02	7.61	5.99	0.04	0.03
MMV Center	39.85	23.44	0.072	0.040

Observers Predicting Other Observer Matches

Clearly there will be variability in the least-dissimilar matches made by the different observers. To what extent do the observers agree with one another and is a match made by an individual observer any better or worse a predictor of the average observer match than those made by the various computational methods?

To answer this question, we used a leave-one-observer-out comparison in which the one observer is excluded and the 9 remaining trials (3 observers, 3 repeats per paper, per illumination condition) are averaged. The mean of the excluded observer's 3 trials is then used as a predictor of the 3-observer average. This process is repeated for each of the four observers resulting in four predictors Obs1,…, Obs4 making predictions of four, 3-observer averages.

Table 4. Observers versus Computational Methods. Similar to the preceding tables but in this case comparing via the Wilcoxon test how well the methods predict the 3-observer averages of dissimilar matches.

Method 1	Method 2	Err1<=Err2	Err2<=Err1	Equal
KSM^2	Obs1	0	11	9
	Obs2	0	16	4
	Obs3	0	11	9
	Obs4	0	13	7
Relit	Obs1	0	3	17
	Obs2	0	9	11
	Obs3	1	5	14
	Obs4	0	8	12
Wpt	Obs1	0	11	9
	Obs2	0	12	8
	Obs3	0	9	11
	Obs4	0	11	9
CAT02	Obs1	0	12	8
	Obs2	0	16	4
	Obs3	0	13	7
	Obs4	0	17	3
Best Linear	Obs1	0	8	12
	Obs2	0	9	11
	Obs3	1	7	12
	Obs4	0	10	10
MMV Center	Obs1	0	20	0
	Obs2	0	20	0
	Obs3	0	19	1
	Obs4	0	19	1

From Table 4 it is clear that human observers predict the 3-observer average better than the other methods since the numbers in the fourth column are much larger than the third column. This conclusion becomes even clearer if we combine the results in the table over the 4 observers. In that case, we find that the relit color signal performance equals the observer performance in 54 (17+11+14+12) out of 80 cases, Best Fit in 45 cases, Wpt in 37 cases, KSM^2 in 29 cases and CAT02 in only 22 cases. In only 2 of the 80 cases are any of the computational methods better than an individual observer.

Discussion

The Logvinenko & Tokunaga [1] asymmetric matching experiment is interesting because it is based on least-dissimilar matching of real papers under real lights. The question it addresses differs from that of corresponding color experiments, which tend to abstract color away from what its purpose might be. Given this different set of experimental data, we have evaluated several color signal prediction methods in terms of how well they correspond to observers' least-dissimilar matching.

Firstly, our analysis shows that observers tend to find the given test paper to be the least-dissimilar match paper. Since there is a forced choice of 1 paper out of 20, this does not mean, however, that observers would always consider that paper to be the least-dissimilar if there were an effectively infinite choice of papers. Secondly, the analysis shows that the average color signal of the observer least-dissimilar matches is a relatively good predictor of the color signal of the test paper under the match illuminant. Thirdly, the computational methods Relit, Best Linear, Wpt and KSM^2 are all quite similar in their effectiveness in predicting the average observer match. CAT02 is considerably less effective. However, none of the methods is as effective as each individual observer in predicting the 3-observer average of the other observers' matches. This implies that all the computational methods studied are failing to capture some important aspect of the observers' least-dissimilar matching strategy.

References

[1] A. Logvinenko and R. Tokunaga, "Colour Constancy as Measured by Least Dissimilar Matching," *Seeing and Perceiving*, vol. 24, no. 5, pp. 407-452, 2011.

[2] D. Brainard, W. Brunt and J. Speigle, "Color constancy in the nearly natural image I Asymmetric matches," *Journal of the Optical Society of America A*, vol. 14, no. 9, pp. 2091-2110, 1997.

[3] V. de Almeida, P. Fiadeiro and S. Nascimento, "Color constancy by asymmetric color matching with real objects in three-dimensional scenes," *Visual Neuroscience*, vol. 21, no. 3, pp. 341-345, 2004.

[4] A. Logvinenko, B. Funt, H. Mirzaei and R. Tokunaga, "Rethinking Colour Constancy," *PloS one*, vol. 10, no. 9, pp. e0135029, 2015.

[5] N. Moroney, M. D. Fairchild, R. W. Hunt, C. Li, M. Luo and T. Newman, " The CIECAM02 colour appearance model," *Proc. Tenth IS&T Color Imaging Conference,* pp. 23-27, 2002.

[6] H. Mirzaei and B. Funt, "Gaussian Illuminants and Reflectances for Colour Signal Prediction," In *Color and Imaging Conference*, *Society for Imaging Science and Technology*, pp. 212-216, 2014.

[7] M. Derhak and R. Berns, "Introducing Wpt (Waypoint): A color equivalency representation for defining a material adjustment transform," *Color Research & Application*, vol. 40, no. 6, pp. 535-549, 2014.

[8] B. Funt, and J. Hao, "Non-diagonal color correction," *Proc. IEEE International Conference on Image Processing*, vol. 1, pp. I-481, 2003.

[9] A. Logvinenko, B. Funt and C. Godau, "Metamer Mismatching," *IEEE Trans. on Image Processing*, Vol. 23, No. 1, pp. 34-43, 2014.

[10] G. Finlayson, M. Drew and B. Funt, "Spectral Sharpening: Sensor Transformations for Improved Color Constancy," *Journal of the Optical Society of America A, vol. 11,* no. 5, pp. 1553-1563, 1994.

[11] J. A. von Kries, "Chromatic Adaptation," *In D. L. MacAdam (Ed.), Sources of Colour Science*. Cambridge, 1970.

Author Biographies

Emitis Roshan is a M.Sc. student at the school of Computing Science, Simon Fraser University. She has a M.Sc. in Industrial Engineering from Sharif University of Technology, Iran, 2014, and a B.Sc. in Computer Engineering from the Isfahan University of Technology, 2010. Her current research spans several areas of color vision.

Brian Funt is Professor of Computing Science at Simon Fraser University where he has been since 1980. He obtained his Ph.D. in Computer Science from the University of British Columbia in 1976. His research focus is on computational approaches to modeling and understanding color. E-mail: funt@sfu.ca