Research Article | Vol. 25, No. 14 | 10 Jul 2017 | OPTICS EXPRESS 16378
Optics EXPRESS

Gamut Volume Index: a color preference metric based on meta-analysis and optimized colour samples

QIANG LIU,[1,2,3,*] ZHENG HUANG,[1] KAIDA XIAO,[2] MICHAEL R. POINTER, [2] STEPHEN WESTLAND,[2] AND M. RONNIER LUO[4]

[1] *School of Printing and Packaging, Wuhan University, Wuhan, 430079, China*
[2]*School of Design, University of Leeds, Leeds, LS29JT, United Kingdom*
[3]*Shen Zhen Research Institute, Wuhan University, Shenzhen, 518000, China*
[4]*State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, Hangzhou, 310058, China*
liuqiang@whu.edu.cn

Abstract: A novel metric named Gamut Volume Index (GVI) is proposed for evaluating the colour preference of lighting. This metric is based on the absolute gamut volume of optimized colour samples. The optimal colour set of the proposed metric was obtained by optimizing the weighted average correlation between the metric predictions and the subjective ratings for 8 psychophysical studies. The performance of 20 typical colour metrics was also investigated, which included colour difference based metrics, gamut based metrics, memory based metrics as well as combined metrics. It was found that the proposed GVI outperformed the existing counterparts, especially for the conditions where correlated colour temperatures differed.

OCIS codes: (330.1690) Color; (330.1715) Color, rendering and metamerism; (330.5020) Perception psychology.

References and links

1. D. Nickerson and C. W. Jerome, "Color rendering of light sources: CIE method of specification and its application," Illum. Eng. **60**, 262–271 (1965).
2. K. W. Houser, M. Wei, A. David, M. R. Krames, and X. S. Shen, "Review of measures for light-source color rendition and considerations for a two-measure system for characterizing color rendition," Opt. Express **21**(8), 10393–10411 (2013).
3. K. Smet, W. R. Ryckaert, M. R. Pointer, G. Deconinck, and P. Hanselaer, "Correlation between color quality metric predictions and visual appreciation of light sources," Opt. Express **19**(9), 8151–8166 (2011).
4. T. Khanh, P. Bodrogi, Q. Vinh, and D. Stojanovic, "Colour preference, naturalness, vividness and colour quality metrics, Part 1: Experiments in a room," Light. Res. Technol., 1477153516643359 (2015)..
5. T. Khanh, P. Bodrogi, Q. Vinh, and D. Stojanovic, "Colour preference, naturalness, vividness and colour quality metrics, Part 2: Experiments in a viewing booth and analysis of the combined dataset," Light. Res. Technol., 1477153516643570 (2016).
6. T. Khanh and P. Bodrogi, "Colour preference, naturalness, vividness and colour quality metrics, Part 3: Experiments with makeup products and analysis of the complete warm white dataset," Light. Res. Technol., 1477153516669558 (2016).
7. S. Jost-Boissard, M. Fontoynont, and J. Blanc-Gonnet, "Perceived lighting quality of LED sources for the presentation of fruit and vegetables," J. Mod. Opt. **56**(13), 1420–1432 (2009).
8. L. Jiang, P. Jin, and P. Lei, "Color discrimination metric based on cone cell sensitivity," Opt. Express **23**(11), A741–A751 (2015).
9. Q. Wang, H. Xu, F. Zhang, and Z. Wang, "Influence of color temperature on comfort and preference for LED indoor lighting," Optik (Stuttg.) **129**, 21–29 (2017).
10. M. Wei and K. W. Houser, "Systematic changes in gamut size affect color preference," Leukos **13**(1), 23–32 (2017).
11. Y. Lin, M. Wei, K. Smet, A. Tsukitani, P. Bodrogi, and T. Q. Khanh, "Colour preference varies with lighting application," Light. Res. Technol., 1477153515611458 (2015).
12. M. Wei, K. W. Houser, G. R. Allen, and W. W. Beers, "Color Preference under LEDs with diminished yellow emission," Leukos **10**(3), 119–131 (2014).
13. M. Islam, R. Dangol, M. Hyvärinen, P. Bhusal, M. Puolakka, and L. Halonen, "User preferences for LED lighting in terms of light spectrum," Light. Res. Technol. **45**(6), 641–665 (2013).

#295997 https://doi.org/10.1364/OE.25.016378
Journal © 2017 Received 16 May 2017; revised 13 Jun 2017; accepted 27 Jun 2017; published 30 Jun 2017

Optics EXPRESS

14. M. Royer, A. Wilkerson, M. Wei, K. Houser, and R. Davis, "Human perceptions of colour rendition vary with average fidelity, average gamut, and gamut shape," Light. Res. Technol., 1477153516663615 (2016).
15. F. L. Schmidt and J. E. Hunter, *Methods of Meta-analysis: Correcting Error and Bias in Research Findings* (Sage Publications, 2014), Chap. 2.
16. S. M. C. Nascimento and O. Masuda, "Best lighting for visual appreciation of artistic paintings--experiments with real paintings and real illumination," J. Opt. Soc. Am. A **31**(4), A214–A219 (2014).
17. N. Narendran and L. Deng, "Color rendering properties of LED light sources," Proc. SPIE **4776**, 61–67 (2002).
18. E. E. Dikel, G. J. Burns, J. A. Veitch, S. Mancini, and G. R. Newsham, "Preferred chromaticity of color-tunable LED lighting," Leukos **10**(2), 101–115 (2014).
19. N. Kakitsuba, "Comfortable indoor lighting conditions evaluated from psychological and physiological responses," Leukos **12**(3), 163–172 (2016).
20. B. C. Park, J. H. Chang, Y. S. Kim, J. W. Jeong, and A. S. Choi, "A study on the subjective response for corrected colour temperature conditions in a specific space," Indoor Built Environ. **19**(6), 623–637 (2010).
21. F. Szabó, R. Kéri, J. Schanda, P. Csuti, and E. Mihálykó-Orbán, "A study of preferred colour rendering of light sources: Home lighting," Light. Res. Technol. **48**(2), 103–125 (2016).
22. J. P. Freyssinier and M. Rea, "A two-metric proposal to specify the color-rendering properties of light sources for retail lighting," Proc. SPIE **7784**, 7784V (2002).
23. M. Rea, L. Deng, and R. Wolsey, "NLPIP Lighting Answers: Light Sources and Color," (Polytechnic Institute, 2004).
24. W. Davis and Y. Ohno, "Color quality scale," Opt. Eng. **49**(3), 033602 (2010).
25. K. Hashimoto, T. Yano, M. Shimizu, and Y. Nayatani, "New method for specifying color rendering properties of light sources based on feeling of contrast," Color Res. Appl. **32**(5), 361–371 (2007).
26. W. A. Thornton, "Color-discrimination index," J. Opt. Soc. Am. **62**(2), 191–194 (1972).
27. S. A. Fotios, "The perception of light sources of different colour properties," Doctor of Philosophy Thesis, UMIST, United Kingdom (1997).
28. W. A. Thornton, "A validation of the color-preference index," J. Illum. Eng. Soc. **4**(1), 48–52 (1974).
29. M. R. Luo, "The quality of light sources," Color. Technol. **127**(2), 75–87 (2011).
30. K. A. Smet, J. Schanda, L. Whitehead, and R. M. Luo, "CRI2012: A proposal for updating the CIE colour rendering index," Light. Res. Technol. **45**(6), 689–709 (2013).
31. A. David, P. T. Fini, K. W. Houser, Y. Ohno, M. P. Royer, K. A. Smet, M. Wei, and L. Whitehead, "Development of the IES method for evaluating the color rendition of light sources," Opt. Express **23**(12), 15888–15906 (2015).
32. K. A. G. Smet, W. R. Ryckaert, M. R. Pointer, G. Deconinck, and P. Hanselaer, "Memory colours and colour quality evaluation of conventional and solid-state lamps," Opt. Express **18**(25), 26229–26244 (2010).
33. S. Jost-Boissard, P. Avouac, and M. Fontoynont, "Assessing the colour quality of LED sources: Naturalness, attractiveness, colourfulness and colour difference," Light. Res. Technol. **47**(7), 769–794 (2014).
34. M. S. Rea and J. P. Freyssinier Nova, "Color rendering: A tale of two metrics," Color Res. Appl. **33**(3), 192–202 (2008).
35. P. van der Burgt and J. van Kemenade, "About color rendition of light sources: The balance between simplicity and accuracy," Color Res. Appl. **35**(2), 85–93 (2010).
36. J. M. Quintero, A. Sudrià, C. E. Hunt, and J. Carreras, "Color rendering map: a graphical metric for assessment of illumination," Opt. Express **20**(5), 4939–4956 (2012).
37. Z. Huang, Q. Liu, S. Westland, M. R. Pointer, M. R. Luo, and K. Xiao, "Light dominates colour preference when correlated colour temperature differs," Light. Res. Technol. (posted 6 June 2017, in press).
38. P. R. Mills, S. C. Tomkins, and L. J. Schlangen, "The effect of high correlated colour temperature office lighting on employee wellbeing and work performance," J. Circadian Rhythms **5**(1), 2 (2007).
39. Q. Liu, X. Wan, J. Liang, Z. Liu, D. Xie, and C. Li, "Neural network approach to a colorimetric value transform based on a large scale spectral dataset," Color. Technol. **133**(1), 73–80 (2017).
40. J. Schanda, *CIE Colorimetry* (Wiley Online Library, 2007), Chap. 3.
41. C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," ACM Trans. Math. Softw. **22**(4), 469–483 (1996).

1. Introduction

Since 1965, the CIE Colour Rendering Index *R*a (CRI) [1] has been used as the standard for assessing the color rendering quality of a light source. Limitations of such a measure have been extensively reported [2–6] and it is widely accepted that a full description of light quality actually includes many different aspects, such as colour fidelity [1], naturalness [7], vividness [4], colour discrimination [8] and colour preference [9].

Among the above aspects, colour preference is undoubtedly considered as a very important dimension, since for general conditions people always pay much attention to the visual appreciation of illuminated scenes.

Optics EXPRESS

To date, many psychophysical experiments have been conducted with the aim of finding a metric to accurately predict colour preference [4–6, 9–13]. However, as stated by many researchers [3, 7, 14], it is very difficult to reach a strong conclusion from a single study because of the lack of statistical robustness. That is, a metric derived from a single experiment with a limited number of light sources and test objects may provide a sound description for the original work, but it should not be expected to have good applicability to other lighting conditions. For instance, Khanh *et al*. recently conducted a series of psychophysical experiments in this topic [4–6]. Their aim was to develop a linearly combined metric of existing measures which could better correspond to the visual appreciation shown by the observers. However, it was found that their fitted metrics always varied with the accumulation of their research data, which highlighted the fact that the conclusion of a single study may depend on the original data and thus have questionable external validity. As far as we believe, that is a crucial reason why no single metric has been agreed to perfectly evaluate colour preference.

In 2011, Smet *et al*. contributed an excellent work [3] which assessed the performance of several typical colour rendition metrics by a meta-analysis of the Spearman correlation coefficients between the metric predictions and the subjective ratings of colour rendition with regard to several psychophysical studies. Since a meta-analysis could estimate the true strength of association among several related works and simultaneously correct for the sampling error or within-study variance [15], the conclusion drawn by such an approach is much more convincing.

However, in that work, Smet *et al*. mainly focused on scenarios of approximately the same CCTs (metameric lighting scenarios) while ignoring the scenarios of different CCTs (multi-CCT scenarios). As we believe, unlike colour fidelity, colour preference should not be restricted by a certain CCT, since in many cases people actually want to choose a favorite light in irrespective of this measure [9, 13, 16–21]. In other words, a metric that performs well for metameric lighting scenarios may perform poorly for multi-CCT scenarios, since it depends on a fixed reference light source and is only valid under certain CCT values.

This study, therefore, is intended to develop a robust metric which could perform well in predicting colour preference, not only in metameric lighting scenarios, but also in multi-CCT scenarios. The meta-analysis method, which was adopted by Smet *et al*. for evaluating the performance of existing metrics, was followed here to develop such a measure with a novel optimizing protocol. Following a brief summary of existing colour quality metrics and the collected psychophysical studies, the details of the proposed GVI are described. The performance of such a metric is comprehensively compared to that of 20 existing measures and the final results show that our proposed metric provides the best performance, both for metameric lighting and multi-CCT conditions. Meanwhile, the rationale of the calculation procedures of the GVI is also discussed in detail.

2. Colour quality metrics

In this study, twenty typical colour rendition metrics were involved, which included the CIE Color Rendering Index (CRI) [1], Gamut Area Index (GAI) [22], Full Spectrum Colour Index (FSCI) [23], Colour Quality Scale (CQS: Qa, Qf, Qp, Qg) [24], Feeling of Contrast Index (FCI) [25], Colour Discrimination Index (CDI) [26], Cone Surface Area (CSA) [27], Color Preference Index (CPI) [28], CRI-CAM02UCS [29], CRI2012 [30], IES TM-30 (Rf and Rg) [31], Memory Colour Rendering Index (MCRI) [32], mean chroma shift of CQS (△C*) advocated by Khanh *et al*. [4–6], the arithmetic mean value of GAI and CRI [3,33] as well as two combined metrics named Colour Quality Index (respectively denoted as CQI1 [4] and CQI2 [5]).

Note that it is not the intent of this paper to describe the details of the above mentioned measures, readers are referred to the cited references as well as Houser's review [2] for further information. It is also worth noting that some of these metrics were not deliberately proposed for assessing colour preference. However, the correlation analysis between these metrics and

Optics EXPRESS

colour preference can be found in the literature [3–7, 33], since a proper metric for colour preference is what is actually needed.

In Houser's review, these metrics were approximately divided into three groups according to their colour rendition intents: colour fidelity, colour preference and colour discrimination [2]. As pointed out by Houser, such classifications are not entirely independent. For instance, FCI is a colour preference metric while CDI is a colour discrimination metric. However, both of these indices are based on a gamut area calculation. In this study, therefore, we have grouped the 20 measures according to their calculation methods, since we supposed that such methods would have stronger correlation with the final prediction performance than that of their colour rendition intents.

2.1 Colour difference based metrics

The colour difference based metrics (CRI [1], CRI-CAM02UCS [29], Qa, Qp and Qf [24], CRI2012 [30], Rf [31], CPI [28] and △C* [4–6]) are exclusively relative measures and most were intended for the characterization of colour fidelity. According to these metrics, the colour difference (or chroma shift) between a set of colour samples under the test source and a reference illuminant of the same CCT are calculated in a certain color space. Generally speaking, the later measures in this group are mainly updated versions of the CIE CRI, with an improved chromatic adaptation transform, a more uniform color space as well as using different colour samples.

2.2 Relative gamut based metrics

The relative gamut based metrics (Qg [24] and Rg [31]) compute the relative gamut area of a set of colour samples under a test light source in a defined colour space. The values of such metrics are normalized by the gamut area of the same colours illuminated by a reference illuminant of the same CCT.

2.3 Absolute gamut based metrics

Unlike relative gamut based metrics, the absolute gamut based metrics (GAI [22], CDI [26], CSA [27] and FCI [25]) do not rely on a reference illuminant but directly calculate the absolute gamut of the colour samples under a test light source. In this group, CSA is not reliant upon a reference illuminant while GAI, CDI and FCI are with constant reference illuminant.

2.4 Memory based metric

The memory based metric MCRI [32] uses colour memory as its reference. Based on empirically derived similarity functions, such a measure evaluates the colour rendition of a light source by comparing the rendered colours of certain familiar objects to their actual memory colours.

2.5 Combined metrics

The combined metrics are the linear combination of a set of current measures. The concept of combining GAI and CRI was firstly raised by Rea *et al*. [34] while the arithmetic mean value between the two measures were used by Smet [3] and Jost-Boissard [33]. The Colour Quality Indices (respectively denoted as CQI1 [4] and CQI2 [5]) were proposed by Khanh *et al*., the CQI1 is a linear combination of CCT and MCRI while CQI2 is a linear combination of CCT, MCRI and △C*.

2.6 Other metric

In addition to the above metrics, there is another metric, FSCI [23], with a different calculation method. It is an absolute measure quantifying the difference between the Spectral Power Distributions (SPD) of a test source and that of an equal-energy spectrum.

Optics EXPRESS

As discussed above, the 20 metrics involved in this study employ different calculation methods. However, apart from FSCI, the calculation of the metrics exclusively relies on a defined set of colour samples. It is widely acknowledged that an appropriate set of samples is of great importance for metric performance [11] and it should cover the entire hue circle in a defined colour space [32, 35]. Meanwhile, there are also other works indicating that certain colours (a saturated red, for instance) seem to be more important than others [7, 13, 14]. Therefore, it may be concluded that the colour samples used in those measures should span the entire hue circle but with some weighting applied. What is more, as pointed out by previous researchers [3, 24], the use of highly-saturated colours may improve the metric performance. As we believe, such an assumption needs to be further investigated, since the colour samples adopted in existing measures are not saturated enough. To our knowledge, no past studies have comprehensively investigated these issues.

In addition, it should be mentioned that there are several complicated methods such as multi-measure approaches and graphical metrics [14, 24, 31, 36]. Such methods may actually exhibit better performance in colour preference evaluation, since they provide much more useful information than a single measure. However, such measures are beyond the scope of this study due to the fact that they are complicated and overwhelming for most naive users in general applications. They are perhaps more suitable for industrial applications.

3. Psychophysical studies

In this work, the experimental data of eight psychophysical studies were collected, which totally contained 16 metameric lighting scenarios and 16 multi-CCT scenarios. The data included the SPDs of the light sources as well as the corresponding subjective ratings of the loosely-defined visual appreciation (*Preference*, *Attractiveness* and *Pleasantness*). For detailed information of these studies, the readers are referred to the cited articles. The following introduction provides an overview of the key points of each study.

3.1 Wei et al. (metameric lighting, 2014)

Eighty-seven participants compared the colour preference of two 3000 K light sources using a 6-point rating method [12]. The paired comparison experiment was implemented within two side-by-side rooms which contained the same coloured objects and still life arrangements. The illumination level at the object location was approximately 250 lx.

3.2 Royer et al. (metameric lighting, 2016)

Twenty-eight observers were asked to rate the colour quality of twenty-six 3500 K light sources in a room filled with several coloured objects [14]. The illumination is approximately 210 lx but not perfectly uniform. An 8-point rating method was used to quantify the observers' judgments for the experimental light sources in terms of several quality descriptors: *Normal-Shift*, *Saturated-Dull* and *Like-Dislike*. In this study, the data for the *Like-Dislike* scale were used.

3.3 Jost-Boissard et al. (metameric lighting, 2009)

In 2009, Jost-Boissard *et al*. investigated the color rendering of fruit and vegetables in terms of *Attractiveness*, *Naturalness* and *Suitability* [7]. Six 3050 K light sources and six 3950 K light sources were used to illuminate four groups of fruit and vegetables (red, green, yellow and multicoloured). A 3050 K halogen light and a 3950 K fluorescent light were respectively adopted as references. The illumination level was approximately 230 lx. A panel of 40 observers participated in the paired comparison experiment and they were asked to choose the appropriate lighting conditions according to their subjective judgment. In this study, the data regarding *Attractiveness* were used. We believe that the experimental objects to some extent influence the colour preference of lighting [11], so we discussed the colour preference of different groups of

Optics EXPRESS

fruit and vegetables separately. Therefore, 8 lighting scenarios (3050 K: red, green, yellow, multicoloured; 3950 K: red, green, yellow, multicoloured) were obtained from this work.

3.4 Jost-Boissard et al. (metameric lighting, 2014)

The second work of Jost-Boissard *et al*. was quite similar [33]. Several aspects of perceived colour quality were investigated by a paired comparison approach (with no fixed reference) from the following aspects: *Naturalness*, *Colourfulness*, *Visual Appreciation* and *Colour Difference*. 45 observers assessed 9 light sources at 3000 K while 36 observers assessed 8 light sources at 4000 K, with an illumination level approximately 220 lx. From this study, the data of 4 scenarios (3000 K: fruits and vegetables, Color Checker chart; 4000 K: fruits and vegetables, Color Checker chart) regarding to *Visual Appreciation* were obtained.

3.5 Szabó et al. (metameric lighting, 2016)

The work of Szabó *et al*. investigated the human preference of home lighting through real-scene (kitchen and living room) experiments [21]. 97 observers, 69 young and 28 elderly, were involved. For the kitchen (CCT = 4000 K) and living room (CCT = 3000 K) scenarios, two groups of different light sources were used and the illumination of the two scenes were both set to 350 lx. For each group of lights, there were 5 SPDs with inconstant FCI values and 5 SPDs with constant FCI values. The subjects were asked to evaluate the colour rendition of those testing lights in term of *Pleasantness*, *Vividness* and *Naturalness*. Since this work is intended to investigate the performance of different metrics in colour preference evaluation, only the SPDs with inconstant FCI values together with the corresponding visual ratings on *Pleasantness* were used. To be consistent with other studies, we only adopted the data of young observers, although the data of the elderly observers were quite similar. To sum up, 2 lighting scenarios were obtained from this work: kitchen-inconstant-FCI-young and living room-inconstant-FCI-young.

3.6 Liu et al. (multi-CCT, 2017)

In previous studies, we had implemented a series of psychophysical experiments with 14 different objects, which included 4 groups of fruit and vegetables similar with the work of Jost-Boissard *et al*., 5 Chinese traditional calligraphies written on papers with different colours, 4 pieces of artwork with different sizes and colour features and a bunch of multicoloured flowers [37]. Certain SPDs with uniformly sampled CCT values ranging from 2500 K to 6500 K were generated using a colour tunable LED while the illumination level was exclusively set to 200 lx. The number of observers in each lighting scenario ranged from 20 to 60 and they were asked to quantify the colour preference using a 7-point rating method or a 5-level ranking method. From these studies, we collected the data of 14 lighting scenarios in total.

3.7 Narendran et al. (multi-CCT, 2002)

In the work of Narendran *et al*., the authors invited 30 observers to participate in a paired comparison experiment as well as a 7-point rating experiment [17]. Seven light sources with the same illuminance level (approximately 200 lx) but different CCTs (ranging from 2600 K to 5000 K) were employed to illuminate a combination of colour objects. The subjects were asked to respond with their visual preference. As the authors stated, the results of the two visual experiments are closely matched. Therefore, in this study only the data of the 7-point rating experiment were adopted.

3.8 Xu et al. (multi-CCT, 2017)

With a forced choice approach, twelve observers were asked to quantify their perceived colour quality for a printed photograph which was illuminated by several different light sources [9]. The independent variables of such a study were 12 CCTs (ranging from 2000 K to 100,000 K), 3 illuminance levels (350 lx, 500 lx and 1000 lx) and 2 assumed lighting scenarios (working

Optics EXPRESS

and relaxing). The dependent variables were the scale values of the observers in terms of *Preference* and *Comfortable*. Since light sources with a CCT value higher than 10,000 K are rare in everyday use, we omitted the data for 25,000 K and 100,000 K. In addition, since the working scenario was highly related to human working performance [38] rather that colour preference, the data regarding that issue was also ignored. Thus, ten SPDs together with the average preference ratings under 3 illuminance levels were adopted. Note that it would be more plausible to use the data of 350 lx to be consistent with other studies, but the average preference ratings of the 3 illuminance conditions were the only data available to us. Fortunately, as pointed out by those authors, the preference estimations of the 3 illuminance levels were quite consistent.

4. Gamut Volume Index

The proposed Gamut Volume Index was developed based on the following assumptions. Firstly, as mentioned above, the colour samples used in a metric should span the entire hue circle but should not be distributed uniformly, since different colour regions may have a different influence on the observers' judgments. Secondly, adopting highly-saturated samples would improve the metric performance. As pointed out in previous papers, a light source may exhibit good performance for non-saturated samples while perform poorly with saturated samples [3], especially for RGB (red-green-blue) white LEDs with strong peaks in their spectra [24]. However, the reverse was found not to be the case [32]. Thirdly, it seems that it is more plausible to quantify the colour gamut with a volume-based algorithm than an area-based algorithm, since a 3D solid reveals much more information than a 2D plane in the same colour space. Fourthly, as discussed above, since colour preference should not be restricted by a certain CCT, an absolute measure independent of a reference is needed.

In our previous work [39], a large scale spectral data set has been built with 8560 uniformly distributed colour samples. In this study, we uniformly divided this data set into 18 sub-groups according to the dominant wavelength [40] of each colour under the D50/2 illuminant/observer condition, which resulted in 15 subgroups with positive dominant wavelengths (around 447 samples in each subgroup) and 3 subgroups with negative dominant wavelengths (around 614 samples in each subgroup). We then selected the most saturated colour sample from each group by excitation purity [40]. Since dominant wavelength and excitation purity respectively correspond to hue and saturation in CIE Colorimetry, we finally got 18 saturated colours which uniformly covered the entire hue circle.

Figure 1 indicates the gamut comparison between the 18 saturated samples and other sample sets of several typical metrics in CIE 1931 xy chromaticity diagram (a) and the a*b* plane of CIELAB colour space (b). It is obvious that the gamut of this sample set is remarkably larger than that of the others, especially in the green and yellow regions.

The newly-built sample set was then weighted by choosing 14 samples from the overall 18 samples. Mathematically speaking, there were $C_{18}^{14} = 3060$ combinations in total. For a certain combination of 14 samples and a certain lighting scenario, a Spearman correlation coefficient was calculated, from the subjective rating scores for the experimental light sources and the corresponding gamut volumes computed in CIELAB colour space with a convex hull algorithm [41]. Since there were 32 lighting scenarios as mentioned above, 32 Spearman correlation coefficients were obtained for each sample combination. Therefore, 3060 combinations resulted in 3060*32 coefficients.

Afterward, we calculated a weighted average Spearman correlation coefficient for each sample combination and obtained 3060 weighted average coefficients. Finally, the 14 samples which corresponded to the maximum value of the 3060 weighted average Spearman correlation coefficients were defined as the optimized colour samples.

Optics EXPRESS

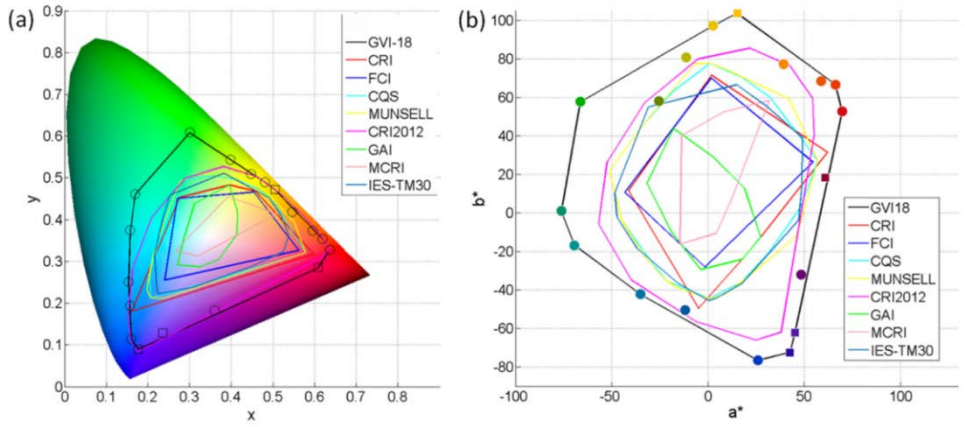

Fig. 1. Gamut comparison of colour sample sets used by different metrics in (a) the xy chromaticity diagram and (b) the a*b* plane under the D50/2 illuminant/observer condition. The circles represent the samples adopted by GVI while the squares represent the omitted samples. The lines denote the gamut boundary derived from a convex hull algorithm.

The method for computing the weighted average correlation coefficient $\bar{r}$ was proposed by Hunter and Schmidt [15]:

$$\bar{r} = \sum_{i=1}^{K} N_i r_i / \sum_{i=1}^{K} N_i \tag{1}$$

where r_i and N_i respectively represent the individual correlation coefficient and the number of observers for each lighting scenarios, K is the number of the scenarios (K = 32 in this study).

The circles and squares in Fig. 1 denote the samples adopted and omitted in the optimized sample set respectively. As can be seen from the figure, three colour samples in blue and purple (including red-purple) regions were omitted, while in other regions only an orange colour was omitted.

In light of the above, the final equation for the GVI could be summarized as:

$$GVI = 5 * V_{optset} \tag{2}$$

where V_{optset} is the gamut volume of the optimized 14 colour samples under the test light source in CIELAB colour space (calculated by a convex hull algorithm [41]) and the constant 5 is used to rescale the metric to an approximate 0-100 range.

5. Results of metric performance analysis

Table 1 summarizes the performance of different measures in terms of the weighted average Spearman correlation coefficient between metric predictions and preference ratings of individual studies. The p-value, which denotes the statistical significance of $H_0 : \bar{r} = 0$ (no correlation), was computed following the equations in the work of Smet *et al*. [3]. The variance of the population correlation σ_ρ^2 indicates the difference among the correlation coefficients of the 32 scenarios with regard to a certain metric and it was calculated by subtracting the sampling error variance σ_e^2 from the variance of the sample correlation σ_r^2 using the following equations as proposed by Hunter and Schmidt [15].

$$\sigma_\rho^2 = \sigma_r^2 - \sigma_e^2 \tag{3}$$

$$\sigma_r^2 = \sum_{i=1}^{K} \left[N_i (r_i - \bar{r})^2 \right] / \sum_{i=1}^{K} N_i \tag{4}$$

Optics EXPRESS

$$\sigma_e^2 = (1-\bar{r}^2)^2 / (\overline{N}-1) \quad (5)$$

where r_i and N_i respectively represent the individual correlation coefficient and the number of observers for each lighting scenarios, K is the number of the scenarios (K = 32 in this study), $\bar{r}$ is the weighted average correlation coefficient described in Eq. (1) and $\overline{N}$ denotes the average value of N_i.

Table 1. The weighted average spearman correlation coefficient between metric predictions and preference ratings of individual studies.

Colour quality metric	metameric lighting			Multi-CCT			Overall performance		
	$\bar{r}$	p	σ_ρ^2	$\bar{r}$	p	σ_ρ^2	$\bar{r}$	p	σ_ρ^2
CRI	-0.15	0.079	0.19	-0.30	0.000	0.07	-0.22	0.000	0.14
CAM02UCS	0.03	0.389	0.16	-0.30	0.000	0.06	-0.12	0.037	0.14
Qa (9.0.3)	0.02	0.416	0.19	-0.30	0.000	0.07	-0.13	0.038	0.16
Qp (9.0.3)	0.60	0.000	0.35	-0.29	0.000	0.06	0.20	0.039	0.41
Qf (9.0.3)	0.17	0.033	0.14	-0.30	0.000	0.06	-0.04	0.271	0.16
CRI2012	0.24	0.018	0.21	-0.08	0.149	0.10	0.09	0.109	0.19
Rf	0.02	0.440	0.17	-0.20	0.005	0.10	-0.08	0.111	0.14
CPI	0.08	0.251	0.23	-0.30	0.000	0.07	-0.09	0.115	0.19
ΔC*	0.67	0.000	0.06	-0.63	0.000	0.10	0.08	0.251	0.48
Qg (7.4)	0.54	0.000	0.35	-0.30	0.000	0.06	0.16	0.073	0.39
Rg	0.85	0.000	0.01	-0.28	0.000	0.07	0.34	0.000	0.34
GAI	0.73	0.000	0.05	0.31	0.000	0.06	0.54	0.000	0.10
CDI	0.73	0.000	0.05	0.31	0.000	0.06	0.54	0.000	0.10
CSA	0.71	0.000	0.06	0.31	0.000	0.06	0.53	0.000	0.10
FCI(CAM02)	0.79	0.000	0.04	-0.30	0.000	0.07	0.30	0.002	0.34
FSCI	0.33	0.017	0.39	0.39	0.000	0.07	0.36	0.000	0.24
MCRI	0.73	0.000	0.03	0.10	0.100	0.10	0.45	0.000	0.16
GAI-CRI	0.65	0.000	0.06	0.45	0.000	0.07	0.56	0.000	0.07
CQI1	0.72	0.000	0.03	0.31	0.000	0.06	0.54	0.000	0.08
CQI2	0.67	0.000	0.06	-0.49	0.000	0.09	0.15	0.092	0.40
GVI	0.85	0.000	0.01	0.81	0.000	0.04	0.83	0.000	0.03

Optics EXPRESS

Figures 2 and 3 further demonstrate the metric performance with regard to each lighting scenario. The values of the correlation coefficients are denoted by colour, for instance, red for very high correlation while blue for very low correlation.

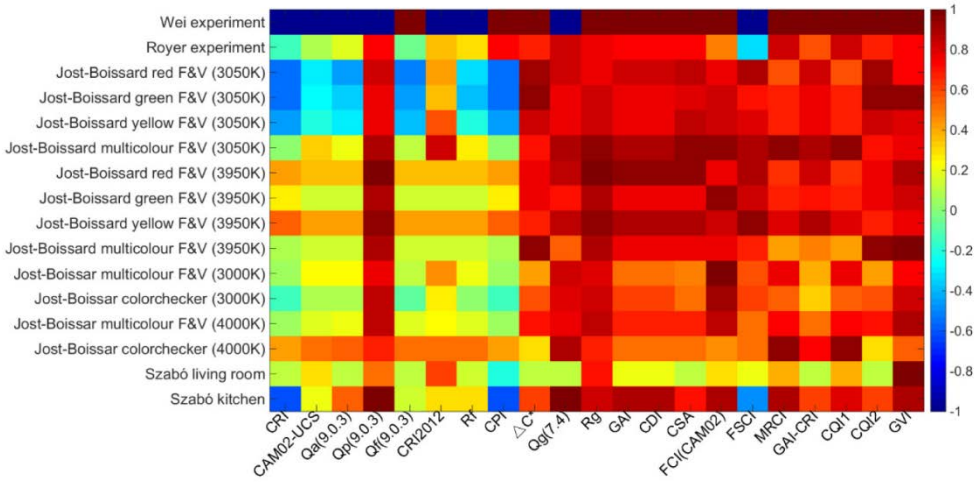

Fig. 2. Weighted average Spearman correlation coefficient between metrics prediction and visual scaling of colour preference of each metameric lighting scenario. ('F & V' is short for 'fruit and vegetables')

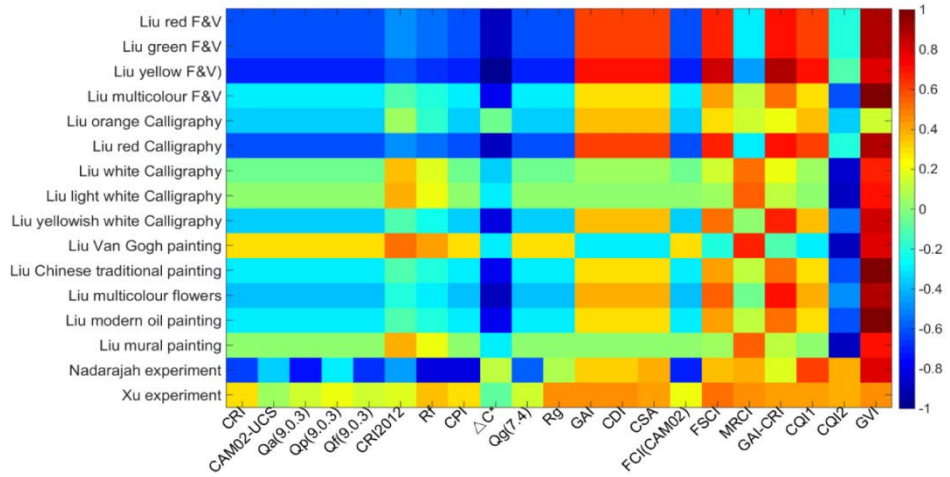

Fig. 3. Weighted average Spearman correlation coefficient between metrics prediction and visual scaling of colour preference of each multi-CCT scenario.

6. Discussion

As can be seen from Table 1, the colour difference based metrics performed poorly in most of the lighting scenarios while CRI had the worst performance among all the measures. The only two exceptions were Qp and △C* for metameric lighting scenarios, which exhibited relatively better correlations. It seems that updating the traditional CRI with stronger theories and models only has very limited effect in improving the metric performance of preference prediction. Such conditions should be attributed to the general intent of colour difference metrics, which is to evaluate colour fidelity, rather than characterize colour preference. According to such measures, only the light sources similar to certain standard references receive high scores. However, as

Optics EXPRESS

pointed out by a previous study [3], there are many light sources that perform better than the standard references. Similarly, another reason for their poor performance in multi-CCT scenarios is that the values of such measures at different CCTs are not comparable, since they are correlated to different reference sources [17, 37].

Compared to the colour difference based metrics, the gamut based metrics generally exhibited better performance. This finding corroborates the thought that colour preference is closely related to chroma (gamut) enhancement [12, 13, 21, 24, 33]. In fact, for metameric lighting scenarios, the sound performance of the above mentioned two colour difference based measures (Qp and △C*) was also due to their rewarding of chroma enhancement. In addition, because of the limitation of the relative measures as mentioned above, the two relative gamut based measures, Qg and Rg performed poorly in multi-CCT scenarios. It is also worth mentioning that even for the absolute gamut measures, the metric performance for the multi-CCT scenarios was not good enough. Such a drawback may possibly be attributed to the distributions of the colours samples, as well as to the methods of computing the gamut area.

The memory based MCRI performed well in metameric lighting scenarios but poorly in multi-CCT scenarios. A possible explanation is that when CCTs differed, the larger deviations among the errors of chromatic adaption transforms impaired the metric performance. That is, for metameric lighting scenarios, the errors of chromatic adaption transforms had almost no impact on the metric computation, since they were approximately consistent. However, in multi-CCT scenarios where those errors significantly differed (especially in highly-saturated regions), their influence became much stronger and could not be ignored. Furthermore, note that in some related works [3, 33] the computation of MCRI was adjusted according to the colour distributions of the experimental objects (i.e., a blue or purple sample was omitted in the case where there was no blue and purple objects in the experiment). Such adjustment was refused in this study since it was intended to find a simple and universal measure for general applications.

Compared to other metrics, the arithmetic mean of GAI and CRI (GAI-CRI) showed a balanced performance between metameric lighting scenarios ($\bar{r}$ = 0.65) and multi-CCT ($\bar{r}$ = 0.45) scenarios, although it was still not good enough. Since CQI1 is a linear combination of CCT and MCRI, in metameric lighting scenarios this measure exhibited similar performance ($\bar{r}$ = 0.72) to that of MCRI ($\bar{r}$ = 0.73), while for multi-CCT scenarios a better performance ($\bar{r}$ = 0.31) was achieved. Note that although such combined metrics actually improved the metric performance, their results were also not good enough. In addition, as shown in Table 1, the performance of CQI2 is obviously worse than that of CQI1, in spite of the fact that CQI2 is based on a larger experimental data set [4, 5].

It is clear from Table 1 that the proposed GVI exhibits significantly better performance, not only for metameric lighting scenarios ($\bar{r}$ = 0.85, $p<0.00001$), but also for multi-CCT scenarios ($\bar{r}$ = 0.81, $p<0.00001$). In addition, the lowest value of variance of population correlation also validated this conclusion.

Figures 2 and 3 straightforwardly confirm the conclusion from Table 1 and reveal more detailed information about the metric performance in each individual lighting scenario. As can be seen from these two figures, the performance of the measures actually varies with the lighting scenarios, which was consistent with the work of Lin *et al*. [11] and highlighted the necessity of discussing each scenario separately. In addition, it is quite clear that these two figures could be considered as good evidence for the drawback of the existing metrics as well as the superiority of GVI, especially for the multi-CCT scenarios.

Note that although the proposed GVI exhibited excellent performance in preference predictions, it is however clear from Figs. 2 and 3 that there were still two exceptions: the orange calligraphy scenario of Liu *et al*. and the experiment of Xu *et al*. For the orange calligraphy scenario, the very low correlation (r = 0.15) was attributed to the fact that the average ratings of three of the experimental lights (4500 K, 5500 K and 6500 K) were almost

Optics EXPRESS

the same. In that condition, it is very likely that the true correlation between the metric prediction and preference scaling was masked by the system errors (for instance, the intra-observer variability) of the experiment. As for the scenario of Xu *et al*., the relatively low correlation (r = 0.45) was due to the over-saturation effect of two lights (8100 K and 9700 K). As stated by current researchers, excessive saturated colours also impaired preference [10]. Among the studies described in Section 3, the GVI values of most of the lights were located in the range of 70-100, while the values for the 8100 K and 9700 K lights of Xu *et al* were above 110.

As shown in Fig. 3, several measures exhibited the same correlation for a certain scenario (e.g., for the 'Liu red F & V' scenario, the 5 colour difference based metrics, CRI, CRI-CAM02UCS, Qa, Qp and Qf shared the same correlation coefficient). Such condition could be ascribed to the dominant influence of light on colour preference under multi-CCT scenarios [37]. Since those measures employed very similar calculation methods, in multi-CCT scenarios they may result in a consistent result in terms of rank order. When calculating the rank-order based Spearman correlation coefficient, therefore, a similar result would be obtained.

To further validate the sample selection method of GVI, the performance of other forms of GVI with different colour sample set was investigated. In Table 2, GVI-CRI-14, GVI-MCRI-10, GVI-CQS-15, GVI-FCI-4, GVI-CRI2012-17, GVI-GAI-8, GVI-IES-99, GVI-MUN-1269 respectively denotes the GVI values which were computed with the colour samples of CRI, MCRI, CQS, FCI, CRI2021, GAI, IES method, as well as the Quintero's graphical metric (Munsell samples) [36], where the number indicates the amount of colour samples in each set. GVI-MUN-14 represents the GVI values with the 14 optimized samples derived from the 1269 Munsell data set by the methods described in Section 4. Similarly, GVI-p08-14, GVI-p07-14, GVI-p06-14 and GVI-p05-14 respectively refer to the GVI values with the 14 optimized samples obtained from the purity-restricted subset of the above mentioned large scale data set with the same optimizing approach. For instance, p08 represents the subset in which the excitation purities of the colour samples were no larger than 0.8. Finally, GVI-uniform-18 denotes the performance of our 18 saturated colour samples without the following weighting implementation and GVI-proposed-14 is our final proposed measure.

From Table 2, several conclusions could be drawn, at least in the conditions of this study. Firstly, the absolute gamut-volume based metrics indeed have advantages. When compared to Table 1, it is clear to see that such measures outperform the 20 existing metrics. Secondly, adopting highly-saturated samples may actually improve the metric performance. For instance, GVI-MUN-14 and GVI-proposed-14 were computed according to a similar approach but with a different original data set and it is quite clear that GVI-proposed-14 with a larger sample gamut performed better. Such a statement was further validated by the case of GVI-p08-14, GVI-p07-14, GVI-p06-14 and GVI-p05-14, where a measure corresponding to a more saturated data set also exhibited better performance. In our opinion, this result may possibly be due to the fact that when saturated samples were used, the diversity of the solid shapes regarding to different sample combinations (under different lights) increased correspondingly, which raised the possibility of obtaining a better GVI. Besides, the Multi-CCT scenarios seem to be more sensitive to the gamut of the test colour samples. For measures with large sample gamut (i.e., GVI-proposed-14 and GVI-CRI2012-17), the performance is good while for measures with small gamut (i.e., GVI-GAI-8 and GVI-MCRI-10), the performance is poor. Thirdly, a weighted distribution of colour samples may also benefit the metric performance. The comparison between GVI-uniform-18 and GVI-proposed-14 is a good example. Although the samples of GVI-uniform-18 were distributed more uniformly, the GVI-proposed-14 performed better, which proved the former assumption that the colour samples used in a metric should span the entire hue circle but should not necessarily be distributed uniformly. In addition, as for the two measures with medium-sized gamut, GVI-CQS-15 and GVI-FCI-4, their good performance may be partially attributed to the reasonable distribution of the colour samples.

Optics EXPRESS

Fourthly, the increase of the number of colour samples provides no benefit to the metric performance, as shown in the cases of GVI-IES-99 and GVI-MUN-1269.

Table 2. The weighted average spearman correlation coefficient between metric predictions and preference ratings of individual studies regarding to other forms of GVI with different sample set.

GVI with different colour samples	metameric lighting			Multi-CCT			Overall performance		
	$\bar{r}$	p	σ_ρ^2	$\bar{r}$	p	σ_ρ^2	$\bar{r}$	p	σ_ρ^2
GVI-CRI-14	0.81	0.000	0.04	0.26	0.000	0.04	0.56	0.000	0.09
GVI-MCRI-10	0.81	0.000	0.02	−0.06	0.233	0.02	0.41	0.000	0.08
GVI-CQS-15	0.73	0.000	0.03	0.64	0.000	0.05	0.69	0.000	0.04
GVI-FCI-4	0.79	0.000	0.02	0.66	0.000	0.04	0.73	0.000	0.04
GVI-CRI2012-17	0.79	0.000	0.01	0.67	0.000	0.08	0.73	0.000	0.05
GVI-GAI-8	0.79	0.000	0.03	−0.27	0.131	0.03	0.30	0.017	0.08
GVI-IES-99	0.78	0.000	0.01	0.26	0.000	0.04	0.54	0.000	0.07
GVI-MUN-1269	0.79	0.000	0.01	0.26	0.000	0.04	0.55	0.000	0.08
GVI-MUN-14	0.75	0.000	0.04	0.65	0.000	0.04	0.71	0.000	0.04
GVI-p08-14	0.79	0.000	0.03	0.76	0.000	0.02	0.78	0.000	0.02
GVI-p07-14	0.81	0.000	0.03	0.75	0.000	0.02	0.77	0.000	0.03
GVI-p06-14	0.81	0.000	0.01	0.53	0.000	0.06	0.68	0.000	0.06
GVI-p05-14	0.79	0.000	0.03	0.35	0.000	0.04	0.59	0.000	0.08
GVI-uniform-18	0.67	0.000	0.02	0.74	0.000	0.08	0.71	0.000	0.05
GVI-proposed-14	0.85	0.000	0.01	0.81	0.000	0.04	0.83	0.000	0.03

Meanwhile, it should be noted that when we built the GVI metric, apart from the optimized 14 colour samples, there are many other sample combinations among the whole 3060 combinations which could achieve good metric performance. To further investigate the colour distributions of those optimal combinations, the best 100 sample combinations with the largest 100 values of weighted average Spearman correlation coefficient were selected. Among those combinations, the worst overall average correlation $\bar{r}$ is 0.77. Figure 4 indicates the colour sample distributions of the best 100 sample combinations. For instance, the ordinate of the #1 colour (dark blue) is 60, which means such a colour was selected 60 times among the best 100 combinations.

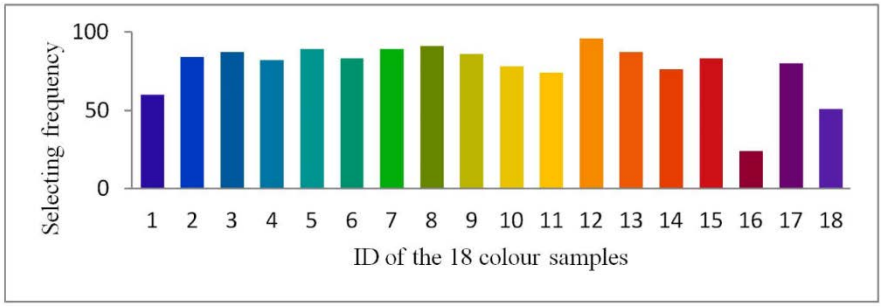

Fig. 4. The colour distributions of the best 100 sample combinations. The Y-axis represents the frequency of a sample (denoted by the colour of the bar) selected in those combinations.

Optics EXPRESS

It can be seen from Fig. 4 that the colours in the pure blue, red-purple, and blue-purple regions (i.e., #1, #16, #18) were less important. Such a finding relates well with the recent work of Royer *et al*. [14], which highlighted the importance of red, orange and green regions while de-emphasized the influence of blue and purple. As we believe, such a condition may be due to the fact that the blue and purple objects do not commonly appear in our daily lives, especially in natural scenes.

7. Conclusions

In this study, an absolute gamut volume based metric (GVI) was developed based on meta-analysis and optimized colour samples. The performance of such a measure was comprehensively compared to 20 typical colour quality metrics, especially in the form of the weighted average correlation between metric predictions and preference ratings of 8 psychophysical studies (32 scenarios). The final results showed that the proposed GVI exhibited the best performance for characterizing colour preference, not only for metameric lighting scenarios, but also for multi-CCT scenarios. It was found that employing certain highly-saturated but non-uniformly distributed samples could actually improve the metric performance while the concept of an absolute gamut-volume metric also has advantage in colour preference prediction.

There should be no doubt that the proposed GVI, which was derived from a meta-analysis based on several existing studies, is much more convincing compared to the measures of single studies. However, such a measure should not be expected to perform well in all situations, since the collected data is still a limiting factor. To further improve the metric performance, a larger set of experimental data should be accumulated. Fortunately, as stated above, there are many other colour sample combinations providing excellent metric performance, which highlights the potential of further optimization. Besides, since this study simultaneously optimized the metric performance of both metameric lighting scenarios and multi-CCT scenarios, certain compromises had to be made. If, however, in some applications only one kind of these scenarios needs to be discussed, a new metric with better performance could be easily obtained by a similar procedure.

Another suggestion for future work concerns setting a limit for rewarding the chroma enhancement in the proposed measure. As mentioned above, excessively saturated samples also impair colour preference. Therefore, such an over-saturated effect should be penalized in an update version of this measure. To investigate such a topic, psychophysical studies with over-saturated light sources should be implemented.

The experimental data of this current research are available upon request.

Funding

National Natural Science Foundation of China (Project No. 61505149, 61405104); the Open Fund of the State Key Laboratory of Pulp and Paper Engineering of China (Project No 201528); Young Talent Project of Wuhan City of China (Project No 2016070204010111).

Acknowledgments

The colleagues who kindly shared the data of their psychophysical experiments are gratefully acknowledged. We also would like to thank Kevin Smet for sharing the codes of the metrics he developed.