Measuring colourfulness in natural images

David Hasler[a] and Sabine Süsstrunk[b]

[a]LOGO GmbH, Steinfurt, Germany
[b]Audiovisual Communication Lab. (LCAV),
Swiss Fed. Inst. of Tech. (EPFL), Lausanne, Switzerland

ABSTRACT

We want to integrate colourfulness in an image quality evaluation framework. This quality framework is meant to evaluate the perceptual impact of a compression algorithm or an error prone communication channel on the quality of an image. The image might go through various enhancement or compression algorithms, resulting in a different—but not necessarily worse—image. In other words, we will measure quality but not fidelity to the original picture.

While modern colour appearance models are able to predict the perception of colourfulness of simple patches on uniform backgrounds, there is no agreement on how to measure the overall colourfulness of a picture of a natural scene. We try to quantify the colourfulness in natural images to perceptually qualify the effect that processing or coding has on colour. We set up a psychophysical category scaling experiment, and ask people to rate images using 7 categories of colourfulness. We then fit a metric to the results, and obtain a correlation of over 90% with the experimental data. The metric is meant to be used real time on video streams. We ignored any issues related to hue in this paper.

Keywords: Image quality metric, colourfulness metric

1. INTRODUCTION

Modern pictorial imaging systems aim at producing the best looking picture rather than at achieving luminance and colour fidelity. While evaluating the quality of a processed image, one needs to consider that if the resulting image is different from the original one, it does not necessarily mean that it is of worse quality. When designing a colour quality metric, we believe that two main factors have to be considered: colour cast and colourfulness. In this paper, we will only consider the overall colourfulness of an image, without measuring fidelity.

We want to quantify 'how bad' is the colour in an image after compression. Our work is part of a larger framework for measuring the perceptual quality of a video stream after transmission over a network, using a *no reference* quality metric approach. The method should be able to work on a single image—or a single video stream—without having the original image. In other words, we cannot determine the quality of a compression and coding scheme by doing an image-based comparison between a compressed image and its original, because the original image is simply not available. Ideally, the method should be able to say if an image is good, but more practically, the scheme might use some meta data that comes along with the data, for example a set of parameters defining the properties of the original image. Additionally, the idea of not using the original image for assessing quality enables the method to deal with images that have gone through various tone mapping or image enhancement algorithms.

Colour can get degraded in two ways: by colour casts or by a colourfulness loss. Modern colour appearance models[1, 2] are able to compute colourfulness correlates of colour patches depending on the viewing conditions and surround. Nevertheless, there is no agreement on how to measure the overall colourfulness of a natural scene, although very recent techniques try to adress *image* colour quality in a more general framework.[3] To try to answer the question of image colourfulness, we set up a psychophysical experiment, where the subject are asked to rate the colourfulness by choosing among 7 categories. Finally, we try to get an algorithm that best fits the result of the psychophysical experiment.

email: david.hasler@bluewin.ch, sabine.susstrunk@epfl.ch

This paper starts by describing the psychophysical experiment (section 2), and the method used to analyse the data (section 3). Following section describes every parameter that is considered for building a metric (section 4), along with the description of the method used to compute an optimal parameter set (section 5). The results are shown next (section 6), followed by a section that might interest anyone concerned with efficient implementations (section 7), where a metric that uses a much simpler colour space is proposed.

2. THE EXPERIMENT

We use 20 non expert viewers and ask them to give a global colourfulness rating for a set of 84 image. The experimental conditions are described in.[4] The user has to choose among the following categories:

1. not colourful
2. slightly colourful
3. moderately colourful
4. averagely colourful
5. quite colourful
6. highly colourful
7. extremely colourful

Prior to the experiment, 4 examples are shown, rated as 'not colourful', 'slightly colourful', 'averagely colourful' and 'extremely colourful' to set the scale of the experiment. None of the examples show the same scene content than the test images. We chose the 2 images in the middle of the scale after conducting a preliminary experiment, using 5 expert viewers, and selecting the images rated with the least confusion among the viewers. The 2 images in the extremity of the scale are chosen by the first author. We used 10 scenes, which we processed by linearly reducing the chroma in CIELab space to generate the 84 test images. The images are shown on a LCD monitor. The images are presented in random order, one image at a time on a grey background. A grey screen lasting $300ms$ is displayed between each image. A subset of the images is shown in figure 1.

We choose to use a category scaling experiment, instead of a paired comparison experiment, to ensure that the viewer adapts to the image white point, and to avoid the influence one image may have on the perception of the other one. Since we consider that a greyscale image has no colourfulness, we can compute a ratio scale using Thurstone's law of comparative judgement, as described in Engeldrum.[5]

3. COMPUTING A SCALE VALUE FROM THE EXPERIMENTAL DATA

We briefly summarise the method found in Engeldrum[5] in section 10.2.2—The reader not interested in implementation issues might as well skip this section. The use of a scale value allows to consider that the perceptual distance between 'slightly colourful' and 'moderately colourful' might be different than the distance between 'highly colourful' and 'extremely colourful'. As we have to attach numbers to these attributes, it is worth trying to get a perceptually uniform scale. For example, if there is a lot of confusion in the judgment between 'slightly colourful' and 'moderately colourful', i.e. a lot of images were rated in both categories by different people while there is almost no confusion in the judgement between 'highly colourful' and 'extremely colourful', this would mean that the distance between 'highly colourful' and 'extremely colourful' is larger than the distance between 'slightly colourful' and 'moderately colourful'.

We will assume that the correlation between the categories as well as the discriminal dispersion of the categories and the samples are constant (by 'samples' we mean the answers of the individual test persons). We start by building a frequency matrix where the elements $\{K_{jg}\}$ are the number of times the image j has been put in category g. We define the cumulative proportion matrix with entries P_{jg} as

$$P_{jg} = \frac{\sum_{k=1}^{g} K_{jk}}{\sum_{k=1}^{m} K_{jk}}$$

where m is the number of categories ($m = 7$). From probability P_{jg} we derive the z-scores z_{jg}. P_{jg} and z_{jg} are related through

$$P_{jg} = \frac{1}{\sqrt{2\pi}} \int_{-z_{jg}}^{\infty} e^{-\frac{1}{2}\omega^2} d\omega.$$

Let t_g be the (unknown) boundary value between the categories, and s_j be the (unknown) scale value for each category. The fundamental assumption underlying the scale computation is that

$$t_g - s_j = z_{jg}. \tag{1}$$

This can be put in matrix form as

$$\mathbf{z} = \mathbf{X} \cdot \mathbf{y} \tag{2}$$

$$\mathbf{y} := [t_1 \ ... \ t_{m-1} \ s_1 \ ... \ s_m]^T \tag{3}$$

where $\mathbf{z}$ is a column vector containing all the z-scores z_{jg}, $\mathbf{X}$ is a matrix used to make (2) equivalent to (1)* and $\mathbf{y}$ is the unknown. If we know $\mathbf{y}$, we know the scale values and the boundaries between the scales. The scale values s define the distances *between* the categories, and thus have an arbitrary absolute value. Consequently, in order to have a solution for (2), we impose an additional contraint, namely that

$$\sum_j s_j = 0,$$

which is implemented by adding a line to matrix $\mathbf{X}$ and appending a 0 to vector $\mathbf{z}$. The whole computation of scale values is based on the fact that there is confusion among the observers. If there are images that get unanimous ratings, they do not provide any scale information, and thus have to be removed from the computation. Finally, the scale values are obtained by solving (2), thus

$$\mathbf{y} = (\mathbf{X}^T\mathbf{X})^{-1} \cdot \mathbf{X}^T\mathbf{z}.$$

4. THE METRICS

To compute a colourfulness metric, we study the distribution of the image pixels in the CIELab colour space.[6] We assume that the image colourfulness can be represented by a linear combination of a subset of the following quantities:

1. σ_a : The standard deviation along the a axis.
2. σ_b : The standard deviation along the b axis.
3. $\sigma_{ab} = \sqrt{\sigma_a^2 + \sigma_b^2}$: The trigonometric length of the standard deviation in ab space.
4. μ_{ab} : The distance of the centre of gravity in ab space to the neutral axis.
5. $A_{ab} = \sigma_a \cdot \sigma_b$: A pseudo-area in ab space.
6. σ_C : The standard deviation of Chroma.
7. μ_C : The mean of Chroma
8. σ_1 : The largest standard deviation in ab space (found by searching the direction in the ab plane along which the standard deviation is maximum).
9. σ_2 : The second largest (i.e. the smallest) standard deviation in ab space.
10. $A_{12} = \sigma_1 \cdot \sigma_2$: the area in ab space.

*$\mathbf{X}$ is composed of 1, 0 or -1 only.

11. σ_S : The standard deviation of Saturation, calculated as Chroma over Lightness.

12. μ_S : The mean of Saturation.

By choosing a subset of these quantities, for example $\{\sigma_a, \sigma_b, \mu_{ab}\}$, we can express the colourfulness of the image using a linear combination of them: $Q = \alpha_1 \cdot \sigma_a + \alpha_2 \cdot \sigma_b + \alpha_3 \cdot \mu_{ab}$. The parameters $\{\alpha_1, \alpha_2, \alpha_3\}$ are found by maximising the correlation between the experimental data and the metric, according to Section 5.

5. COMPUTING THE METRIC PARAMETERS

We want to obtain the parameter vector α ($\alpha := [\alpha_1 \cdots \alpha_m]^T$) that correlates the most with the experimental data. To get a meaningful analysis—one that can be generalised to other images—it is important not to use the same data in computing the correlation and in optimising the parameter set. One possibility is to use half of our N images to compute the correlation, and the other half optimise the parameter set. Since the number of images is quite small, we will compute the optimal parameter set using $N-1$ images, and use it to compute the colourfulness of the remaining image. We will repeat this experiment N times, to obtain N colourfulness values that are used to compute the correlation of the metric with the experimental data.

Let $\hat{M}_i$ be the colourfulness computed from image i. By assuming that we are using a subset of m parameters $\mathbf{x}^{(i)} := \left[x_1^{(i)} \cdots x_m^{(i)}\right]^T$ of image i among the parameters of Section 4, the colourfulness can be expressed as

$$\hat{M}_i = \alpha^T \mathbf{x}_i.$$

The parameters $\{\alpha_j\}$ are found by maximising the correlation between the other $N-1$ images of the test set and the experimental values M^{exp} found through the subjective testing:

$$\{\alpha_j\}_i = \underset{\alpha_2 \cdots \alpha_m}{\arg\max} \frac{\sum_{k \neq i} (\tilde{C}_k - \mu_{\tilde{C}}) \cdot (M_k^{\text{exp}} - \mu_{M^{\text{exp}}})}{\sqrt{\sum_{k \neq i} (\tilde{C}_k - \mu_{\tilde{C}})^2 \cdot \sum_{k \neq i} (M_k^{\text{exp}} - \mu_{M^{\text{exp}}})^2}} \tag{4}$$

$$\alpha_{1,i} := 1,$$

$$\tilde{C}_k := \sum_{j=1}^{m} \alpha_j \cdot x_j^{(k)}, \tag{5}$$

where $\mu_\cdot$ denotes the mean value of $(\cdot)$. Since the parameter vector α is defined up to a constant factor, we set arbitrarily $\alpha_1 := 1$.

The correlation ρ between the experimental data and the metric is found using

$$\rho = \frac{\sum_{k=1}^{N} (\hat{M}_k - \mu_{\hat{M}}) \cdot (M_k^{\text{exp}} - \mu_{M^{\text{exp}}})}{\sqrt{\sum_{k=1}^{N} (\hat{M}_k - \mu_{\hat{M}})^2 \cdot \sum_{k=1}^{N} (M_k^{\text{exp}} - \mu_{M^{\text{exp}}})^2}} \tag{6}$$

Finally, the optimal parameter vector α is found by taking the mean value of the N parameter sets defined in (4).

$$\alpha = \frac{1}{N} \sum_{k=1}^{N} [\alpha_1 \cdots \alpha_m]_k^T.$$

Instead of this value, we also could have taken the parameter set that maximises the correlation between the experimental data and the metric using *all* images. Note that the variance of parameters α_i gives an indication of how stable the optimal parameter set is with respect to the choice of the images.

Parameter subset	correlation	metric details
$\sigma_1, \sigma_2, \mu_C$	94.2%	$\sigma_1 + 1.46 \cdot \sigma_2 + 1.34 \cdot \mu_C$
$\sigma_a, \sigma_b, \mu_{ab}$	94.0%	$\sigma_a + \sigma_b + 0.39 \cdot \mu_{ab}$
σ_{ab}, μ_C	94.0%	$\sigma_{ab} + 0.94 \cdot \mu_C$
σ_{ab}, μ_{ab}	93.7%	$\sigma_{ab} + 0.37 \cdot \mu_{ab}$
$\sigma_a, \sigma_b, \mu_C$	93.6%	$\sigma_a + 0.78 \cdot \sigma_b + 0.72 \cdot \mu_C$
$\sigma_1, \sigma_2, \mu_{ab}$	93.5%	$\sigma_1 + 0.81 \cdot \sigma_2 + 0.43 \cdot \mu_{ab}$
σ_S, μ_S	92.3%	$\sigma_S + 1.6 \cdot \mu_S$
σ_C, μ_C	92.1%	$\mu_C + 1.17 \cdot \mu_C$
A_{ab}, μ_{ab}	88.8%	$A_{ab} + 7.3 \cdot \mu_{ab}$
A_{12}, μ_{ab}	87.1%	$A_{12} + 9.3 \cdot \mu_{ab}$

Table 1. Correlation of various colourfulness metrics with the experimental data. Each line corresponds to a different metric, detailed in the last column. The exact formulation has been obtained by an optimisation on the correlation value.

6. RESULTS

By choosing different subset of the attribute described in Section 4, we can try to find the best correlate to the image colourfulness. Table 1 summarises the results. The result range from 94% down to 87% of correlation. To select the best metric, we have to consider several aspects: The most obvious is the correlation to the experiment. The second is the computational cost, and the last is related to the limitation of the experiment due to our initial choice in the selection of the 10 scenes. Provided that the CIELab space has been designed to be a uniform colour space, it does not seem reasonable to emphasize the red-green axis over the blue-yellow axis. The optimisation showing a preference for one of the two axis may be biased by the choice of the test images. In other words, we prefer the parameter σ_{ab} to a sum of σ_a and σ_b, also because σ_{ab} does not depend on the arbitrary direction of the a and b axis. For computational reasons, we avoid using σ_1 and σ_2 because they require a Singular Value Decomposition (SVD), without delivering substantially better results. We also want to avoid using saturation (σ_S and μ_S), since it over-emphasises dark areas, precisely the area that get very roughly approximated by compression algorithms. Unfortunately, we did not include compressed images in the test set, explaining the good performance of these parameters[†]. Finally, we propose two metrics:

$$\hat{M}^{(1)} = \sigma_{ab} + 0.37 \cdot \mu_{ab} \quad (7)$$

$$\hat{M}^{(2)} = \sigma_{ab} + 0.94 \cdot \mu_C, \quad (8)$$

where each symbol is defined in Section 4. Our colourfulness metric is a linear combination of the mean and standard deviation of the pixel cloud in the colour plane of CIELab. The $\hat{M}^{(1)}$ metric seems more natural, because it is a truly two-dimensional metric. It is also computationally more efficient but has a slightly worse correlation, if we consider that a 0.3% difference in correlation is a significant difference.

7. A MORE EFFICIENT METRIC

In this section, we will try to reproduce the results of Section 6 using a computationally more efficient approach. We use a very simple opponent colour space:

$$rg = R - G$$

$$yb = \frac{1}{2}(R + G) - B$$

[†]We knew from past experiences that saturation is not a good correlate when using compressed images, so we discarded its use beforehand, but finally included it for comparison purposes.[7] The use of compressed images in the test set would probably have confirmed this argument.

Attribute	$M^{(1)}$	$M^{(2)}$	$M^{(3)}$
not colourful	0	0	0
slightly colourful	6	8	15
moderately colourful	13	18	33
averagely colourful	19	25	45
quite colourful	24	32	59
highly colourful	32	43	82
extremely colourful	42	54	109

Table 2. Correspondence between the colourfulness metric, and the colourfulness attributes.

We assume that the image is coded in the sRGB colour space. By reconducting the experiment described in section 5, we get a new colourfulness metric

$$\begin{aligned} \hat{M}^{(3)} &= \sigma_{rgyb} + 0.3 \cdot \mu_{rgyb}, \\ \sigma_{rgyb} &:= \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2}, \\ \mu_{rgyb} &:= \sqrt{\mu_{rg}^2 + \mu_{yb}^2}, \end{aligned}$$

where $\sigma_.$ and $\mu_.$ are the standard deviation and the mean value of the pixel cloud along direction $(\cdot)$, respectively. Surprisingly, the correlation of $\hat{M}^{(3)}$ with the experimental data is equal to 95.3%, thus it represents a very nice and efficient way of computing the colourfulness.

8. HOW TO USE THE METRIC

The metric can be used to determine how colourfulness evolves by passing through a tone mapping or a coding algorithm in the following ways:

$$\Delta M_\varepsilon = \hat{M}_p - \hat{M}_o, \tag{9}$$

$$\Delta M_\% = \frac{\hat{M}_p}{\hat{M}_o}, \tag{10}$$

where $\hat{M}_o$ is the colourfulness estimate of the original image, and $\hat{M}_p$ is the colourfulness estimate of the processed image. We would recommend the use of ΔM_ε over $\Delta M_\%$, but further experimentation would be necessary to confirm this argument.

To give some intuition about the metric, Table 2 summarises the 'meaning' of the metric. For example, a value of $\hat{M}^{(3)} = 59$ means that the images is quite colourful.

9. CONCLUSIONS

We tried to introduce colour in an image quality metric scheme, and found that measuring colourfulness was a very promising way to achieve this goal. We set up a psychophysical experiment and asked the viewers to rate the colourfulness of an image picturing a natural scene. We then studied several metrics using the CIELab colour space, and found a simple metric which correlates to about 94% with the experimental data. We also proposed another metric, which is very easy to compute, and achieves an even better correlation (95%) to the experimental data. This metric can be used to evaluate the performance of a coding scheme in real time.

We did not consider hue in our experiments. Nevertheless, a complete colour metric should take hue into account, for example by measuring colour casts between the original and the processed image.

10. ACKNOWLEDGMENTS

We want to thank Genista Corp. for sponsoring this research. We also want to thank the Audiovisual Communication Lab. at EPFL for allowing us to use their Laboratory for the experimental tests. Finally, we want to thank all the viewers that took part in the testing.

APPENDIX A. DEFINITION OF THE IMAGE ATTRIBUTES

This section briefly defines the parameters used in Section 4

Let I_p be the pixel values of an image in Lab space, $p = 1 \cdots N$. The image has N pixels.

$$\begin{aligned}
I_p &:= [L_p \ a_p \ b_p]^T \\
\sigma_a^2 &:= \frac{1}{N}\sum_{p=1}^{N}\left(a_p^2 - \mu_a^2\right) \\
\mu_a &:= \frac{1}{N}\sum_{p=1}^{N} a_p \\
\mu_{ab} &:= \sqrt{\mu_a^2 + \mu_b^2} \\
C_p &:= \sqrt{a^2 + b^2} \\
\mu_C &:= \frac{1}{N}\sum_{p=1}^{N} C_p \\
\sigma_C^2 &:= \frac{1}{N}\sum_{p=1}^{N}\left(C_p^2 - \mu_C^2\right) \\
S_p &:= \frac{C_p}{L_p} \\
\mu_S &:= \frac{1}{N}\sum_{p=1}^{N} S_p.
\end{aligned}$$

The parameters σ_1 and σ_2 need a Singular Value Decomposition (SVD) computation. Let $\mathbf{U}$ and $\mathbf{V}$ be two orthogonal matrices. Let $\mathbf{I}$ be the matrix containing the colour of all the pixels of the image.

$$\mathbf{I} := \begin{bmatrix} a_1 & \cdots & a_N \\ b_1 & \cdots & b_N \end{bmatrix}^T.$$

The matrix $\mathbf{I}$ can be written as

$$\mathbf{I} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T,$$

where $\mathbf{S}$ is a diagonal matrix. Finally σ_1 and σ_2 are computed as

$$[\sigma_1 \ \sigma_2] = [\sigma_a \ \sigma_b] \cdot \mathbf{V}^T$$

REFERENCES

1. N. Moroney, M. D. Fairchild, R. R. Hunt, C. Li, R. M. Luo, and T. Newman, "The CIECAM02 color appearance model," in *IS&T/SID Tenth Color Imaging Conference*, 2002.
2. C. Li, R. M. Luo, R. R. Hunt, N. Moroney, M. D. Fairchild, and T. Newman, "The performance of CIECAM02," in *IS&T/SID Tenth Color Imaging Conference*, 2002.

3. M. D. Fairchild and G. M. Johnson, "Meet iCAM: A next-generation color appearance model," in *IS&T/SID Tenth Color Imaging Conference*, 2002.
4. S. Winkler and R. Campos, "Video quality evaluation for internet streaming applications," in *Proceedings of IS&T/SPIE: Human Vision and Electronic Imaging VIII, IS&T/SPIE*, **5007**, (Sant Clara, CA, USA), 2003.
5. P. G. Engeldrum, *Psychometric Scaling: A Toolkit for Imaging Systems Development*, Imcotek Press, 2000.
6. R. Hunt, *Measuring Colour*, Fountain Press, England, 3 ed., 1998.
7. S. Yendrikhovskij, F. Blommaert, and H. de Ridder, "Perceptually optimal color reproduction," in *Proceedings of SPIE: Human Vision and Electronic Imaging III*, **3299**, pp. 274–281, (San Jose, CA, USA), 1998.

(a)

(b)

(c.2)

(c)

(d)

Figure 1. Images used in the experiment. (a),(b) and (c) are used in the scaling experiment. (d) is shown as example before the experiment. (c.2) is has been obtained from (c) by linearly reducing the chroma in Lab space—the blue/purple colour shift that arises in the operation should not affect the results since we are comparing images that are different from each other. (a),(b) and (c) are taken from the Corbis royalty free collection.