Learning Image Similarity Measures from Choice Data

Matthias Scheller Lichtenauer (1)(3), Peter Zolliker (1), Ingmar Lissner (2), Jens Preiss (2), Philipp Urban (2);
(1) Laboratory for Media Technology, Empa, Duebendorf, Switzerland, (2) Institute of Printing Science and Technology, Technische Universität Darmstadt, Germany, (3) Friedrich-Schiller-Universität, Jena, Germany

Abstract

We present a corpus of experimental data from psychometric studies on gamut mapping and demonstrate its use to develop image similarity measures. We investigate whether similarity measures based on luminance (SSIM) can be improved when features based on chroma and hue are added.
Image similarity measures can be applied to automatically select a good image from a sample of transformed images.

Keywords: image quality, gamut mapping, psychometry

Introduction

Image transformation algorithms such as gamut mapping can be adapted to a particular image. But the choice of good parameters is often far from trivial even for experienced creative professionals. If more than two parameters have to be chosen, an extensive manual trial–and–error strategy becomes cumbersome. So, there is a need for an automatic method to qualitatively assess the outcome of an image transformation with respect to the original. We will derive such a method from human choices.
The effects of lossy compression or transmission on images have been studied extensively and reference databases exist [1–5]. In this context, it arises naturally to derive features from differences of two images. But image difference features can also be applied to assess image transformations other than compression or transmission. Wang et al. [6, 7] introduced the term of structural similarity (SSIM) for a perceptual difference measure based on luminance features of two compared images. Barańczuk et al. [8] predicted the perceived quality of gamut mapping based on image difference features including structural similarity.
There is a large body of gamut mapping studies using the method of paired comparisons, hence categorical choice information only, while most of the databases cited above use mean opinion scores on difference scales. The law of comparative judgement or conjoint analysis can be used to derive values on a difference scale from choice data [9–12]. Such a scale value could be useful to make studies based on paired comparison comparable with studies reporting mean opinion scores. But if choice shall be predicted, one might also directly design classifiers without deriving scale values first.
In this paper, we evaluate image difference features and similarity measures in the context of gamut mapping. Our ground truth is based on paired comparison experiments. Each experiment involved various mappings of an image to target gamuts.
The remaining paper is organised as follows. First, we present a corpus of study data we plan to make publicly available. Then, we apply these data in order to develop an image difference measure. We will compare the indirect approach — deriving a metric score in order to predict choice based on this score — with a direct classification based on machine learning methods. Results are summarised and discussed prior to our conclusions.

A database of gamut mapping studies

The studies included in this database were all conducted between 2006 and 2011. All studies used the experimental method of paired comparison. Table 1 contains a summary of the number of images involved and the number of trials (comparisons). For each study, we recorded

- the original and the transformed images,
- records of choices with anonymised observer ID,
- derived score data for each mapped image,
- a description of the test setup.

At the moment of writing this, we have included twelve studies with a total of more than 70'000 choices.
In the first four studies, there was only one parameter identifying the algorithm used [8, 13–15]. The studies denoted by 'Mixing' segmented images and applied one out of a candidate set of algorithms for each segment. Different schemes on how to fuse segments back to an image were compared [16]. In the conjoint studies, a couple of parameters per transformation algorithm were recorded, including four target gamut sizes [11].

Table 1. A summary of the studies

Name	images	trials	non-tied	ties	%
Basic	97	5550	5199	351	6.3%
Local Contrast	72	5376	5209	167	3.1%
Image Gamut	65	4087	3698	389	9.5%
Individual	20	8000	8000	0	0.0%
Mixing 1	36	4327	3900	427	9.8%
Mixing 2	36	4816	3659	1157	24.0%
Mixing 3	36	5400	4713	687	12.7%
Mixing 4	50	5320	4869	451	8.5%
Mixing 5	36	4664	3739	925	19.8%
Mixing 6	50	6036	5123	913	15.1%
Conjoint 1	85	3186	2860	326	10.2%
Conjoint 2	95	13068	11401	1667	12.7%

Derivation of scores from choice data

We include scores for each transformed image in the database. Scores are not only used to apply regression models predicting choice for images that are not in the database, but we provide them also to facilitate comparative studies with image quality databases that use mean opinion scores (MOS) on difference scales.
All included studies displayed an original and two transformed images on a computer screen in each trial. Observers were instructed to either abstain (no abstention possible in Individual study) if both transformed images seemed equal or else to click on the image which seemed to be the better representation of the original.

The method to derive a score varied depending on the number of choices per stimulus combination:

- **Global score**: When several choices were recorded for each combination of stimuli, a model based on Thurstone's case V can be applied to derive a score [9, 10, 12].

- **Mixed score**: When the number of choices per stimulus combination was small, mixed regression was applied. For this, scores based on global performance of an algorithm were mixed with individualised scores based on comparisons involving a particular original. Cross validation methods were used to determine the mixing proportions [8].

- **Conjoint case**: In the conjoint case (multiple parameters per algorithm), a linear model of independent attributes was used to derive a score, again combined with mixing global and individualised scores [11].

In the database, we provide the global score, the score based on each original only and the score based on the mixture of both.

Application of the database

We used the database to develop an image similarity measure derived not only from luminance features, but also from features involving chroma and hue.
As a benchmark for the quality of the image similarity measure, we did not use prediction of scores, as these are already derived data. Instead, we used the percentage of correctly predicted observer choices (hit rate). For this, we excluded some choices as a test set and derived a predictor using the other choices in a subset of the database as a training set. If not declared otherwise, both subsets included only data taken from the same study. The choices in the test set had to be predicted. Repetition of this cross validation process generated statistics of hit rate for a particular predictor. The hit rate reported here is defined as the number of correct predictions divided by the number of non-tied choices. Although it would have been negligible, the denominator was decreased by the number of tied predictions.
We will now describe the calculation of the features and then present two methods used as predictors of choice.

Calculation of structural similarity features

The SSIM measure [7] is defined for one channel only, typically a luminance channel, and consists of three basic features, one responsible for average luminance differences, and the other two for contrast and structural differences.

Let x indicate a window around a pixel position in the original image X, and let y be a corresponding window in the mapped image Y. The SSIM formula as we use it reads as follows:

$$\mathrm{SSIM}(X,Y) = \overline{\mathrm{l}(x,y)}^{\alpha_1} \cdot \overline{\mathrm{c}(x,y)}^{\alpha_2} \cdot \overline{\mathrm{s}(x,y)}^{\alpha_3} \quad , \tag{1}$$

with overlines indicating averaging over all windows and the following luminance feature functions per window:

$$\mathrm{l}(x,y) = \frac{(2\mu_x\mu_y + c_1)}{(\mu_x^2 + \mu_y^2 + c_1)} \quad , \tag{2a}$$

$$\mathrm{c}(x,y) = \frac{(2\sigma_x\sigma_y + c_2)}{(\sigma_x^2 + \sigma_y^2 + c_2)} \quad , \tag{2b}$$

$$\mathrm{s}(x,y) = \frac{(\sigma_{xy} + c_3)}{(\sigma_x\sigma_y + c_3)} \quad , \tag{2c}$$

where $\alpha_1 > 0$, $\alpha_2 > 0$ and $\alpha_3 > 0$ are parameters used to adjust the relative importance of the three components. The constants are set to the values in [7] here, namely $c_1 = (0.01L)^2$, $c_2 = (0.03L)^2$ and $c_3 = \frac{1}{2}c_2$. The symbols μ_x and σ_x denote empirical mean and standard deviation in the sliding window, σ_{xy} is the correlation between corresponding windows. L is the numeric dynamic range of the pixel values e.g. 255 for 8-bit RGB and 100 for CIELAB, respectively.
Formally, SSIM is defined on an intensity. This does not have to be the luminance, but could also be chroma – but not hue.

However, we also want to have the possibility to consider colour related similarities, particularly hue. Therefore, we extend the SSIM measure in equation (1) by additional features for hue and chroma shifts:

$$\chi\mathrm{SSIM}(X,Y) = \mathrm{SSIM}(X,Y) \cdot \overline{\chi(x,y)}^{\alpha_4} \cdot \overline{\mathrm{h}(x,y)}^{\alpha_5} \quad . \tag{3}$$

The two additional features for chroma and hue are set up with $c_i > 0$ from the following feature functions per window:

$$\chi(x,y) = \frac{1}{c_4 \cdot \Delta\mathrm{C}(x,y)^2 + 1} \quad , \tag{4}$$

$$\mathrm{h}(x,y) = \frac{1}{c_5 \cdot \Delta\mathrm{H}(x,y)^2 + 1} \quad . \tag{5}$$

The hue and chroma differences ΔH and ΔC can be calculated from Cartesian chromaticity coordinates a and b as follows:

$$\Delta\mathrm{C}(x,y) = \sqrt{a_x^2 + b_x^2} - \sqrt{a_y^2 + b_y^2} \quad , \tag{6}$$

$$\Delta\mathrm{H}(x,y) = \sqrt{(a_x - a_y)^2 + (b_x - b_y)^2 - \Delta\mathrm{C}(x,y)^2} \quad . \tag{7}$$

The feature functions $\overline{f(x,y)}$ fulfil the criteria set by [7], namely

1. Symmetry: $\overline{f(x,y)} = \overline{f(y,x)}$;
2. Boundedness: $\overline{|f(x,y)|} \leq 1$;
3. Unique maximum: $\overline{f(x,y)} = 1$ if and only if $x = y$ (in discrete representations, $x_i = y_i$ for all i = 1; 2; ... N);

We do not consider structural correlations in chroma and hue since the perception of structural information is mainly governed by luminance. In order to derive a meaningful hue measure, a hue preserving colour space has to be used. Here, we use IPT [17] and DIN6164 [18, 19] as working colour spaces, which are known to have good hue preserving properties. Hence, we apply the standard SSIM formula to the luminance coordinate in the respective working colour spaces.

Method 1: Linear regression on scores

As a first method to predict choices, we applied linear regression with score as dependent variable to different subsets of features such as the algorithm ID, three components of SSIM, as well as features based on chroma and hue described in the last paragraph (see Table 2) as independent variables. the chromatic features were calculated in both, IPT and DIN colour spaces [17–19].

We restrict the optimisation here to finding optimal exponents α_i based on the data in our database. We can transform the problem into a linear system by applying the logarithm to equation (3) :

$$\alpha_i \cdot \log\left(\frac{1}{c_i \cdot f(x,y)^2 + 1}\right) \approx -\alpha_i \cdot (c_i \cdot f(x,y)^2) \quad , \tag{8}$$

where $f(x,y)$ denotes chroma or hue differences described above. Note that exponents α_4 and α_5 are strongly correlated with the parameters c_4 and c_5.
Thus, to a good approximation, a change in α_i by a factor can be compensated by a change in c_i by the inverse factor. Without changing the ordering properties of the similarity measure, we can scale all exponents in equation (3) by an arbitrary factor. In order to be comparable to the original SSIM, we set the average of the first three exponents to 1. Furthermore, we adapt the parameters c_4 and c_5 after linear regression such that the exponents α_4 and α_5 become 1 when using equation (8).

Table 2. Subsets of feature terms used

Subset	Feature terms								
	ID	SSIM RGB	SSIM IPT			SSIM DIN		
		$\bar{l},\bar{c},\bar{s}$	$\bar{l},\bar{c},\bar{s}$	$\bar{\chi}$	$\bar{h}$	$\bar{l},\bar{c},\bar{s}$	$\bar{\chi}$	$\bar{h}$
1	x							
2		xxx						
3			xxx					
4			xxx	x	x			
5			xxx	x				
6			xxx		x			
7						xxx		
8						xxx	x	x
9						xxx	x	
10						xxx		x

Method 2: Support vector classification

Choice can be expressed as classification. We used classifying support vector machines as predictors of choice [20], more specifically the publicy available implementation of LIBSVM in version 3.0 [21].
Support vector machines for binary classification need training data vectors $\vec{x}_1 \ldots \vec{x}_n$ from input data space $X \subseteq \mathbb{R}^d$ as well as their class labels $y_1 \ldots y_n$, $y_i \in \{-1,+1\}$.
Support vector machines find a hyperplane (a high-dimensional equivalent to a plane), separating a space into two half-spaces which should each contain only data of the same class. From all possible separating hyperplanes, the one is selected that maximises the distance (margin) to the nearest points (support vectors). The optimisation problem formulated this way is convex and has therefore only global optima.
The key point is that the distance to the optimal hyperplane can be expressed using inner products of mapping functions $\Phi(\vec{x})$ of the input data. These inner products of mapping functions are implicitly calculated by Mercer kernel functions: $k(\vec{x}_i,\vec{x}) = \langle \Phi(\vec{x}_i), \Phi(\vec{x}) \rangle$. A vector $\vec{x}$ is labelled $+1$ if $b + \Sigma_{i=1}^{n} \alpha_i y_i k(\vec{x}_i,\vec{x}) \geq 0$, else $\vec{x}$ is labelled -1. Here, we used the linear kernel $k(\vec{x}_i,\vec{x}) = \langle \vec{x}_i,\vec{x} \rangle$. Parameters b and $\vec{\alpha}$ are determined in the optimisation process.
Data points on the wrong side of the separating hyperplane can be allowed to extend the model, but they have to be penalised. Since calculating $\Phi(\vec{x}_i)$ is the same as mapping the data to a feature space, which can have different dimensionality and geometry, the separating plane in feature space corresponds in general to a bent surface in input space. A regularisation parameter C is introduced, which balances between minimising curvature of the separating surface and penalisation due to misclassifications. This way, over-fitting the data can be avoided. In order to find a robust regularisation parameter C, the data are split into two partitions, named training and test set. For a particular parameter setting $C = c$, $c \in \mathbb{R}$, $c > 0$, the support vector machine learns a model on the training partition (i.e. finds a separating hyperplane). The learned model is then used to predict the choice on the other, the test partition. The cross validation process is repeated with different partitionings, thus creating statistics on predicting accuracy for a particular value of the parameter C. We chose the $c \in \{2^k : k \in \mathbb{Z}, -4 < k < 4\}$ with highest median accuracy of predicting choices in the test sets.

Prediction of choice

Both methods, linear regression and classifying support vector machines were applied to predict choices based on different subsets of feature terms listed in Table 2. The individual terms were calculated for both transformed images in a trial relative to the original image. While scores were predicted for each transformed image separately by regression models, the respective data of both images were concatenated in order of presentation from left to right so as to form the data vector $\vec{x}$ in the input space of the support vector machines. For each subset of feature terms, a separate machine was trained.

To predict choice from a score, let r_{ai}, r_{bi} be the respective scores for two transformed images in trial i, regardless of whether calculated from a set of choices or regressed with image difference features. The choice predicted from scores is then defined as $y_{ip} = \text{sign}(r_{bi} - r_{ai})$, $y_{ip} \in \{-1,0,+1\}$.
The choice being predicted by support vector machine classification is identical to the predicted class label $y_{ip} \in \{-1,+1\}$.

Results and discussion

As a quality criterion for both prediction methods we used the accuracy (hit rate) in predicting choice on data that were not used in learning or regression (cross validation method). Please note that the maximal achievable hit rate is usually not 100%, as contradicting choices are to be expected in the data.
Hit rates for different feature sets should only be compared if the same data have been used in learning. This was one of the reasons why we decided not to split data randomly for the results in Table 3, but in such a way that data from exactly one image were excluded in each cross validation run. As choice tasks were repeated, we could also ensure this way that the choice to be predicted was not already contained in the training set.
When comparing mean hit rates in Table 3, adding the chroma and hue information almost always improved the mean hit rate compared to those feature sets including only luminance-based information. The observation holds for both colour spaces we used, IPT and DIN.
As a reference hit rate, we include the hit rates when using scores calculated without cross validation on *all* data of each study (last three columns of Table 3). Note that the global score uses only information on the algorithms we used, while in the image individualised and mixed scores, information about images is contained. Feature set 1—the algorithm ID for the non-conjoint studies and the ID of the target gamut for the conjoint studies—does not contain information about individual images, either. For the predictions with regression or support vector machines in Table 3, information about algorithms as well as images that were involved was indirectly contained in the image difference features. Therefore, it is possible to exceed the hit rate achievable when using the algorithm ID or the global score alone.

Table 3. Mean hit rates in cross validation for chosen feature sets compared to Thurstone on *all* choices

	Regression IPT		Regression DIN		SVM IPT		SVM DIN		Thurstone / Conjoint		
Study	Set 3	Set 4	Set 7	Set 8	Set 3	Set 4	Set 7	Set 8	global	image	mixed
Basic	72.7%	73.8%	72.6%	74.1%	72.2%	73.3%	72.8%	73.4%	71.9%	83.4%	81.7%
Local Contrast	66.3%	67.9%	66.5%	68.2%	66.3%	67.8%	65.6%	66.4%	68.0%	79.4%	79.3%
Image Gamut	68.5%	69.3%	68.3%	68.6%	67.9%	68.0%	66.9%	67.2%	71.2%	79.3%	77.9%
Individual	62.2%	62.0%	60.3%	61.6%	60.6%	60.6%	60.4%	60.4%	61.1%	69.8%	69.8%
Mixing 1	64.3%	63.1%	64.6%	65.6%	61.8%	66.1%	62.8%	64.6%	64.5%	73.8%	73.7%
Mixing 2	66.7%	69.2%	66.9%	70.2%	66.1%	70.4%	65.7%	69.4%	71.5%	65.5%	74.5%
Mixing 3	55.7%	57.8%	57.1%	60.0%	54.7%	62.8%	53.8%	59.1%	61.1%	71.5%	71.5%
Mixing 4	60.4%	60.5%	61.0%	63.0%	61.1%	62.1%	62.2%	64.4%	60.8%	72.7%	72.4%
Mixing 5	58.4%	58.8%	58.1%	60.1%	60.8%	61.6%	59.4%	61.4%	60.0%	65.2%	64.0%
Mixing 6	57.4%	56.1%	57.3%	56.7%	53.8%	57.1%	56.3%	58.2%	57.3%	67.7%	67.3%
Conjoint 1	77.7%	81.0%	78.2%	80.8%	79.3%	81.6%	76.7%	81.0%	78.6%	82.0%	82.1%
Conjoint 2	84.9%	86.8%	85.0%	86.7%	85.2%	87.4%	84.4%	87.2%	86.2%	86.7%	87.6%

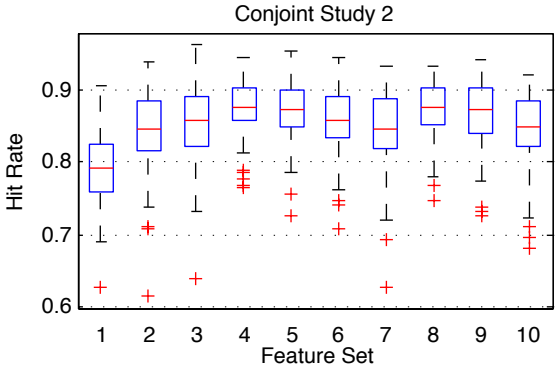


Figure 1. *Hit rate distributions of the support vector machine for each of the features sets in Table 2, data from each original image excluded once as test set. Feature set 2 corresponds to standard SSIM.*

High variability of hit rates between original images (Figure 1) makes it tricky to derive whether the differences in hit rates are significant. Furthermore, the number of comparisons per image is not always identical, not even within one study.

Significance tests with random partitions

We restrict our significance testing to the linear model here, factorial and hybrid combination of features are discussed in [**?**, 22]. We used random partitions of equal size. All predictors compared used the same randomly created training and test sets. Due to repetition of trials, identical trials to the one of which choice should be predicted could be in the training set. Should this influence the hit rate, all predictors would share this information.

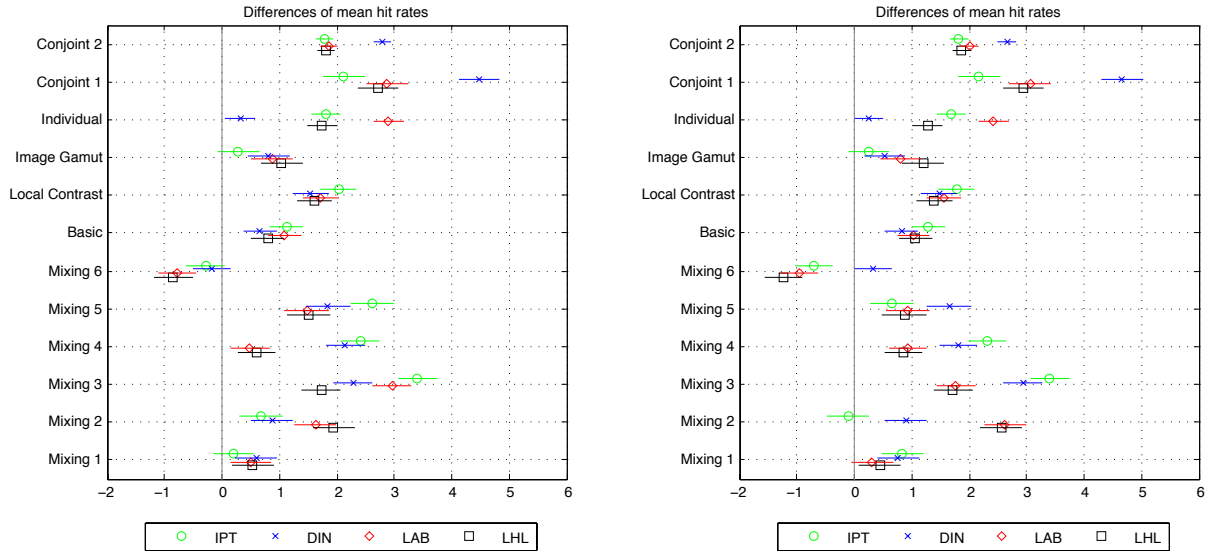
The c_i values in the feature functions used by SSIM were held constant for the significance tests, and the average of all sliding windows of the respective feature function's values was used as a feature. We learned the coefficients of a linear combination of feature differences by ordinary linear least squares regression on randomly selected 90% of the choice data for each study, using each image-difference-based feature combination listed in Table 2 as independent variables. With learned coefficients, we then predicted choice on the remaining 10% of the data. We repeated this cross validation process 1000 times. MATLAB's `anova1` and `multcompare` were then used to test the hypothesis that the mean hit rates of the different feature sets are equal at a confidence level of 99%.

We were particularly interested in the differences in hit rates between regression using only luminance-based features $(\bar{l},\bar{c},\bar{s})$ and regression using luminance, chroma-based and hue-based features together $(\bar{l},\bar{c},\bar{s},\overline{\chi},\bar{h})$. Figure 2 shows the mean of the difference in hit rates between these two feature sets. Positive values mean adding chroma-based and hue-based features performs better in this linear model. The error bars show the 99% confidence interval for differences of means based on the 1000 repetitions for each study. If zero is included in this interval, the difference is insignificant at this level.

We used four working colour spaces for the significance tests. The working colour space abbreviated to LHL was recently presented as LAB2000HL (Hue Linear) by Lissner and Urban [22]. Please note that the gamut mapping operations have not been performed in each of these spaces; we only calculate the difference features from given mapped images represented in these spaces. So, our results should not be used to assess the performance of any of these spaces as working colour space for *mapping* images to a target gamut.

First, we used regression on Thurstone's scale values mixed from global and individualised values as dependent variable [8]. In this case, the features based on luminance, chroma and hue are significantly better for most studies than those based on luminance only, with the exception of Mixing 1 and 6, see Figure 2(a). We repeated the same cross validation of linear regression, but used choice expressed as +1 or -1 as dependent variable. In this setup, the general picture is similar, see Figure 2(b). The hit rates were comparable to those achieved with regression on Thurstone's scale values.

The use of choice predictions from scores or scale values needs a short discussion here: how close to the original an image is after gamut mapping does not only depend on the algorithm, but also on the target gamut size relative to the image gamut size. In most cases, an image mapped from sRGB to an offset gamut will look better than an image mapped to a newspaper gamut, as was experimentally established in the conjoint studies [11]. So, in order to compare Thurstone scale values or mean opinion scores from two studies, there should be a normalisation procedure. As long as there are no identical comparisons contained in both studies, such a normalisation needs a model on its own. In other words, without further knowledge, a difference scale should be considered as valid within one study only, while hit rates on choices may be accumulated.



(a) Linear regression on mixed Thurstone scale value

(b) Linear regression on choice (+1/-1)

Figure 2. *Comparing hit rates of luminance-based feature sets with hit rates of feature sets based on luminance, chroma and hue. If values are positive, the latter perform better.*

Achievable hit rates

We used hit rate as an indicator of the accuracy of a model. Given that observers may make contradicting choices, we wondered what the best accuracy could be.

When one ignores that different images were compared and considers only the information about the algorithms involved, all choices for each algorithm pair can be accumulated as votes. The best (maximal) hit rate given only information about algorithms can be achieved by predicting for each algorithm pair the choice the majority of observers has made. We included the corresponding hit rate in Table 4. We also calculated the hit rate that is achievable if the majority procedure is executed for each original image separately.

Table 4. Mean hit rates achievable (majority hit rates)

	Majority per		Trials per
Study	algorithm	image pair	image pair
Basic	72.4%	98.6%	0.97
Local Contrast	68.0%	81.8%	2.58
Image Gamut	72.6%	84.7%	2.03
Individual	61.2%	70.1%	40.00
Mixing 1	64.5%	76.0%	7.22
Mixing 2	71.5%	77.0%	10.16
Mixing 3	61.9%	73.7%	8.72
Mixing 4	60.8%	79.0%	3.48
Mixing 5	60.3%	69.4%	6.92
Mixing 6	57.4%	74.9%	3.65
Conjoint 1	95.2%	99.8%	< 0.01
Conjoint 2	99.9%	99.9%	< 0.01

For the conjoint studies, there were 1536 possible combinations of algorithm parameters, each counted as a different algorithm. The number of algorithms matters, since the number n of possible image pairs is linear in the number o of originals, but quadratic in the number a of algorithms: $n = o\frac{1}{2}a(a-1)$. One should therefore also consider the average number of trials per possible pair of transformed images (leftmost column in Table 4). Values below 1 in that column mean that most possible combinations were never compared (or the number of ties matters in this context, e.g. in the Basic study). If single comparisons are frequent, very few contradictions can occur at all and the theoretical hit rate approaches 100%. One workaround could be to count only those pairs where more than one comparison was made. But in the conjoint studies there were less than one hundred multiple comparisons out of thousands left in that case–not enough for a reliable estimation of the maximal achievable hit rate. For the other studies, counting only multiple comparisons made a difference of more than 1% only in the Basic study: the maximal achievable hit rate in the Basic study dropped from 98.6% to 84.0%.

Optimisation of the similarity measure

As a last result, we calculate hit rates when optimising parameters of a linear model across all studies in the database. Linear regressions of the proposed extended similarity measure in different working colour spaces (WCS) are shown in Table 5. For the results in Table 5, all non-tied choices of *all studies together* were used and regression was performed on scores mixed within each study as described above.

To be precise, we estimated α_i by a linearised version of $\chi SSIM$, assuming that most features f would be near 1, so that $f^{\alpha_i} = (1-\Delta_f)^{\alpha_i}$. We then used the approximation $\log((1-\Delta_f)^{\alpha_i}) \approx -\alpha_i \cdot \Delta_f$ to estimate the values of α_i. For all hue preserving colour spaces, we see a significant increase in the hit rate of $2-3\%$ when adding features based on hue and chroma to those based on luminance only.

A separate optimisation of the three exponents α_1, α_2 and α_3 increases the hit rate only marginally; thus, with the current data we can set them all to 1. The performance of the different work-

ing colour spaces is similar with no clear preference. This means that SSIM features already that were calculated could be used and simply extended with the chroma-based and hue-based features proposed here.

Influence of experts and lay observers

It is worth noting that algorithm mixture studies 3 and 5 as well as 4 and 6 are paired studies, once with experts in a laboratory setup (3,4), but also under unknown conditions on the Internet (5,6). The differences in favour of using the hue and chroma features are always more significant for the laboratory setting. Possible explanations are differences in criteria between experts and lay people or less homogeneous viewing conditions.

Table 5. Influence of keeping some coefficients fixed on hit rates for χSSIM in different working colour spaces (WCS). Empty entries mean that the corresponding feature was left out.

Keeping the mean of luminance-based coefficients at 1						
WCS	α_1	α_2	α_3	c_4	c_5	hit rate
RGB	1.29	0.98	0.72			66.4 %
LAB	1.27	0.91	0.83			66.4 %
LAB	1.26	0.93	0.81		0.005	66.9 %
LAB	1.12	0.79	1.09	0.01		67.5 %
LAB	1.12	0.8	1.08	0.01	0.01	67.6 %
IPT	1.37	0.89	0.74			66.4 %
IPT	1.37	0.94	0.68		0.01	67.9 %
IPT	1.21	0.77	1.02	0.008		67.4 %
IPT	1.22	0.82	0.97	0.008	0.008	68.3 %
DIN	1.14	0.96	0.9			66.4 %
DIN	1.14	1.01	0.86		0.012	67.1 %
DIN	1	0.84	1.15	0.013		67.2 %
DIN	1	0.88	1.12	0.013	0.013	67.7 %
LHL	1.23	0.92	0.85			66.5 %
LHL	1.23	0.94	0.83		0.005	66.8 %
LHL	1.09	0.79	1.12	0.01		67.6 %
LHL	1.08	0.81	1.11	0.01	0.01	67.7 %
Setting luminance-based coefficients to 1 as in [7]						
WCS	α_1	α_2	α_3	c_4	c_5	hit rate
RGB	1	1	1			66 %
LAB	1	1	1			65.9 %
LAB	1	1	1		0.005	66.2 %
LAB	1	1	1	0.009		67.4 %
LAB	1	1	1	0.009	0.009	67.5 %
IPT	1	1	1			65.7 %
IPT	1	1	1		0.009	67 %
IPT	1	1	1	0.008		67.2 %
IPT	1	1	1	0.008	0.008	68.2 %
DIN	1	1	1			66.3 %
DIN	1	1	1		0.01	67.1 %
DIN	1	1	1	0.011		67.1 %
DIN	1	1	1	0.011	0.011	67.5 %
LHL	1	1	1			66.1 %
LHL	1	1	1		0.004	66.5 %
LHL	1	1	1	0.009		67.5 %
LHL	1	1	1	0.009	0.009	67.6 %

Résumé, conclusions and future work

We presented a database containing choice-based experiments on mapping images of natural scenes to target gamuts.

Accuracy of predicting choices (hit rate) was proposed as a quantitative criterion for the quality of an image similarity measure. Maximal achievable hit rate due to contradicting choices was discussed.

We presented an extension of SSIM with chroma and hue components. This extended image similarity measure significantly improved mean hit rates relative to SSIM based on luminance features alone when evaluated on the presented database.

We conclude that in the context of gamut mapping, features based on chroma and hue improve image similarity measurements based on luminance only. We see our contribution as a first step towards inclusion of colour features into a similarity measure. The underlying hypotheses about the human visual system are the subject of [**?**].

There is still room for optimisation, such as the choice of working colour space for image similarity measures and the best terms to model the perception of hue and chroma differences. A further goal is the verification of the performance of a common similarity measure for different applications such as image compression, transmission and gamut mapping. The use of multiple scales should be evaluated as well. Data for these purposes are available.

We are interested to include more psychovisual studies of other researchers in our database and plan to make it publicly available. In particular, studies are needed which include a greater variety of chroma and hue differences within the compared images in order to derive more reliable parameters for the proposed candidates of the structural similarity terms.

Acknowledgements

Part of this research was supported by the Swiss National Science Foundation.

References

[1] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics. *Advances of Modern Radioelectronics*, 10:30–45, 2009.

[2] H.R. Sheikh, M.F. Sabir, and A.C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006.

[3] P. Le Callet and F. Autrusseau. Subjective quality assessment IRCCyN/IVC database, 2005. http://www.irccyn.ec-nantes.fr/ivcdb/.

[4] Y. Horita, K. Shibata, and Y. Kawayoka. Toyama image quality evaluation database. http:// mict.eng.u-toyama.ac.jp/mict/index2.html.

[5] E.C. Larson and D.M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, 2010.

[6] Z. Wang and A. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, 2002.

[7] Z. Wang, A. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–611, 2004.

[8] Z. Barańczuk, P. Zolliker, and J. Giesen. Image-individualized gamut mapping algorithms. *Journal of Imaging Science and Technology*, 54(3):953–957, 2010.

[9] L.L. Thurstone. A Law of Comparative Judgment. *Psychology Review*, 34:273–86, 1927.

[10] P.G. Engeldrum. *Psychometric Scaling: A Toolkit for Imaging Systems Development*. Imcotek Press, Winchester, MA, USA, 2000.

[11] P. Zolliker, Z. Barańczuk, I. Sprow, and J. Giesen. Conjoint analysis for evaluating parameterized gamut mapping algorithms. *IEEE Transactions on Image Processing*, 19(3):758–769, 2010.

[12] P. Zolliker and Z. Barańczuk. Error estimation of paired comparison tests for Thurstone's Case V. In *5th European Conference on Colour in Graphics, Imaging and Vision*, pages 39–44. IS&T (CD-ROM), 2010.

[13] P. Zolliker and K. Simon. Retaining local image information in gamut mapping algorithms. *IEEE Transactions on Image Processing*, 16(3):664–672, 2007.

[14] J. Giesen, E. Schuberth, K. Simon, and P. Zolliker. Image-dependent gamut mapping as optimization problem. *IEEE Transactions on Image Processing*, 16(10):2401–2410, 2007.

[15] F. Dugay, I. Farup, and Y.I. Hardeberg. Perceptual evaluation of color gamut mapping algorithms. *Color Research and Application*, 33(6):470–476, 2008.

[16] P. Zolliker, Z. Barańczuk, and J. Giesen. Image Fusion for optimizing Gamut Mapping. In *Proceedings of the 19th Color Imaging Conference*, pages 109–114. IS&T, 2011.

[17] F. Ebner and M. D. Fairchild. Development and Testing of a Color Space (IPT) with Improved Hue Uniformity. In *Proceedings of the 6th Color Imaging Conference*, pages 8–13, 1998.

[18] *DIN-Farbenkarte DIN 6164*. Beuth, Berlin und Köln, 1960.

[19] M. Richter and K. Witt. The story of the DIN color system. *Color Res. Appl.*, 11:138–145, 1986.

[20] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[21] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at *http://www.csie.ntu.edu.tw/∼cjlin/libsvm*.

[22] I. Lissner and P. Urban. Toward a unified color space for perception-based image processing. *IEEE Transactions on Image Processing*, 21(3):1153–1168, 2012.

[23] J. Preiss, I. Lissner, P. Urban, M. Scheller Lichtenauer, and P. Zolliker. The impact of image-difference features on perceived image differences. In *6th European Conference on Colour in Graphics, Imaging, and Vision (accepted)*, 2012.

Author Biographies

Matthias SCHELLER LICHTENAUER studied Computer Science at ETH Zürich, receiving his Master of Science in 2008. He then joined the Laboratory of Media Technology at Empa where he is researching the subject of design and analysis of psychometric measurements. He is also a Ph.D. candidate in the group of Joachim Giesen at the Friedrich-Schiller-University in Jena (Germany).

Peter ZOLLIKER studied Physics at ETH Zürich and received his Ph.D. in Crystallography from the University of Geneva in 1987. After his postdoc position at the Brookhaven National Laboratory in New York, he joined Gretag Imaging in 1988. Since 2003 he is working at Empa where he is engaged in colour management and statistical analysis.

Ingmar LISSNER received his degree in Computer Science and Engineering from the Hamburg University of Technology (Germany) in 2009. He is currently working toward the Ph.D. degree with the Institute of Printing Science and Technology, Technische Universität Darmstadt (Germany). His research interests include colour perception, uniform colour spaces, and image-difference measures for colour images.

Jens PREISS received his diploma in Physics (equivalent to a M.S.) from the University of Freiburg (Germany) in 2010. He is currently a research assistant at the Institute of Printing Science and Technology, Technische Universität Darmstadt (Germany), where he works as a doctoral candidate in the area of colour and imaging science.

Philipp URBAN has been head of an Emmy-Noether research group at the Technische Universität Darmstadt (Germany) since 2009. His research focuses on colour science and spectral imaging. From 2006-2008 he was a visiting scientist at the RIT Munsell Color Science Laboratory. He holds a MS in mathematics from the University of Hamburg and a Ph.D. from the Hamburg University of Technology (Germany).