

Perceptual Evaluation of Color-to-Grayscale Image Conversions

M. Čadík^{†1}

¹Czech Technical University in Prague, Czech Republic

Abstract

Color images often have to be converted to grayscale for reproduction, artistic purposes, or for subsequent processing. Methods performing the conversion of color images to grayscale aim to retain as much information about the original color image as possible, while simultaneously producing perceptually plausible grayscale results. Recently, many methods of conversion have been proposed, but their performance has not yet been assessed. Therefore, the strengths and weaknesses of color-to-grayscale conversions are not known. In this paper, we present the results of two subjective experiments in which a total of 24 color images were converted to grayscale using seven state-of-the-art conversions and evaluated by 119 human subjects using a paired comparison paradigm. We surveyed nearly 20000 human responses and used them to evaluate the accuracy and preference of the color-to-grayscale conversions. To the best of our knowledge, the study presented in this paper is the first perceptual evaluation of color-to-grayscale conversions. Besides exposing the strengths and weaknesses of the researched methods, the aim of the study is to attain a deeper understanding of the examined field, which can accelerate the progress of color-to-grayscale conversion.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Display algorithms, viewing algorithms; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Color, shading, shadowing, and texture; I.4.3 [Image Processing and Computer Vision]: Enhancement—Filtering; J.4 [Social and Behavioral Sciences]: Psychology

1. Introduction

Converting color images to grayscale is used for various reasons, like for reproducing on monochrome devices, subsequent processing, or for aesthetic intents. Color-to-grayscale conversions perform a reduction of the three-dimensional color data into a single dimension, seen in Figure 1. It is evident that some loss of information during the conversion is inevitable, so the goal is to save as much information from the original color image as possible. At the same time, the aim is also to produce perceptually plausible grayscale results. Recently, various approaches to the color to grayscale conversion problem have been proposed. While the problem's complexity is currently recognized, the performance of existing solutions is not. Even though researchers frequently claim that their methods advance the field with re-

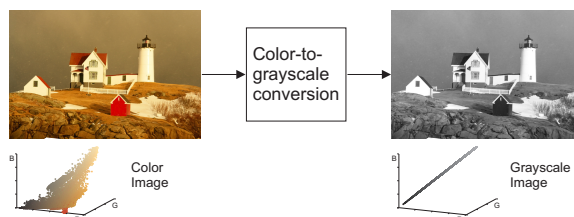


Figure 1: The color to grayscale image conversion.

spect to previous ones, it is important to evaluate the performance of these algorithms in comparative, subjective experiments and analyze their strengths and weaknesses. However, until now, there has not been an evaluation of color-to-grayscale conversions involving a representative number of subjects and input stimuli.

In this paper, we present the results of two subjective perceptual experiments (preference and accuracy), for which

[†] cadikm@fel.cvut.cz <http://www.cgg.cvut.cz/~cadikm>

seven state-of-the-art color-to-grayscale conversions were evaluated by 119 human subjects. The set of inputs consisted of 24 various color images. By means of statistical analysis of the subjective experimental data, we assess the strengths and weaknesses of the conversions, with respect to the preference and accuracy of color reproduction. The overall results show that the best score for *accuracy* is achieved by the approach of Smith et al. [SLJT08], while the most *preferred* method is Decolorize [GD05]. The method of Bala and Eschbach [BE04] was ranked the worst in both the accuracy and preference experiments. Furthermore, we aim to attain a deeper insight into the color-to-grayscale conversion field.

The rest of this paper is structured as follows: In Section 2 we first survey the related work. In Section 3 we introduce the two perceptual experiments that we have conducted. In Section 4 we present, analyze and discuss the results of the experiments. Finally, in Section 5 we conclude and suggest some ideas for future research.

2. Related Work

In this section, we give an overview of current state-of-the-art, color-to-grayscale conversions. Most of the described methods are evaluated in our perceptual study (please, refer to Section 3.1 and Table 1). We also survey existing evaluations of color-to-gray conversions and related studies.

2.1. Color-to-Grayscale Image Conversions

The simplest and widely used approach to converting color to grayscale is based on neglecting of the chrominance channels, e.g. taking a luminance channel as a grayscale representation of the original color image. One of the possibilities is to utilize the Y channel of the CIE XYZ [Fai05] color space. This approach is simple and computationally efficient, but it may fail for specific images, such as those with isoluminant colors.

Bala and Eschbach [BE04] propose a spatial approach to color-to-grayscale conversion. They preserve chrominance edges locally by introducing high-frequency chrominance information into the luminance channel. A spatial high-pass filter is applied to the chromatic channels, the output is weighted with a luminance-dependent term, and the final result is added to the luminance channel.

Grundland and Dodgson [GD05] propose the *Decolorize* algorithm for contrast enhancement as well as converting color to grayscale. They perform a global grayscale conversion by expressing grayscale as a continuous, image-dependent, piecewise linear mapping of the primary RGB colors and their saturation. Three parameters are used to control contrast enhancement, scale selection and noise suppression, and image-independent default values for these parameters have been proposed [GD05].

A different approach was taken by Gooch et al. [GOTG05], who introduced the local algorithm known as *Color2Gray*. In this gradient-domain method, the gray value of each pixel is iteratively adjusted to minimize an objective function, which is based on local contrasts between all the pixel pairs. The computational complexity of this method is high ($O(N^4)$), and can be improved by limiting the number of considered differences (e.g. by color quantization). Mantiuk et al. [MMS06] show an application of their contrast processing framework to accelerate the *Color2Gray* [GOTG05] method. In their approach, the close neighborhood of a pixel is considered on fine levels of a pyramid, whereas the far neighborhood is covered on coarser levels. The authors claim that this enables them to convert bigger images and perform computations faster.

Another conversion was introduced by Rasche et al. [RGW05]. Their method aims to preserve contrast while maintaining consistent luminance. The authors formulate an error-function based on matching the gray differences to the corresponding color differences. The goal is minimizing the error function to find an optimal conversion. The authors propose using color quantization to reduce the considerable computational costs of the error-minimization procedure.

Queiroz and Braun [dQB06] have proposed an *invertible* conversion to grayscale. The idea is to transform colors into high-frequency textures that are applied onto the gray image and can be later decoded back to color. The method is based on wavelet transformations and on the replacement of subbands by chrominance planes.

Alsam and Kolas [AK06] introduced a conversion method that aims to create sharp grayscale from the original color rather than enhancing the separation between colors. The approach resembles the method of Bala and Eschbach [BE04]: first, a grayscale image is created by a global mapping to the image-dependent gray axis. Then, the grayscale image is enhanced by a correction mask in a way similar to unsharp masking [GW02].

Neumann et al. [NČN07] proposed two local, gradient-based, color-to-grayscale conversions. The first is a generalization of the CIELab formula [Fai05], which introduces a signed power function to give a signum to the weighted Lab components. The second technique aims to obtain the best perceptual gray gradient equivalent by exploiting the Coloroid system and its experimental background. The gradient field constructed using one of the techniques is corrected using a gradient inconsistency correction method. Finally, a 2D integration yields the grayscale image.

A recent method by Smith et al. [SLJT08] combines global and local conversions in a way similar to Alsam and Kolas [AK06]. The method first applies global “absolute” mapping based on the Helmholtz-Kohlrausch effect, and then locally enhances chrominance edges using adaptively-weighted multiscale unsharp masking. While the global

conversion	reference	G/L	implementation	parameters
CIE Y	[Fai05]	G	own implementation, C++	—
Bala04	[BE04]	G + L	own implementation, C++	N=3, K=1, B1=15, B2=40
Decolorize	[GD05]	G	www.eyemaginary.com , Matlab	effect=0.5, scale=25, noise= 10^{-3}
Color2Gray	[GOTG05]	L	www.color2gray.info , command_line, C++	colors=256, $\theta=45$, $\alpha=10$, μ =full
Rasche05	[RGW05]	L	www.fx.clemson.edu/~rkarl , c2g_i, C	colors=256, exp=2, threshold=15
Neumann07	[NČN07]	L	www.cgg.cvut.cz/~cadikm/color_to_gray , own impl., C++	$\epsilon = 10^{-5}$
Smith08	[SLJT08]	G + L	www.mpi-inf.mpg.de/resources/ApparentGreyscale , 1-scale, Gimp	rad=5, amount=0.15, gamma=1

Table 1: Summary of the evaluated color-to-grayscale conversion methods. G and L stands for global and local, respectively.

mapping is image independent, the local enhancement reintroduces lost discontinuities only in regions that insufficiently represent the original chromatic contrast [SLJT08]. The goal of the method is perceptual accuracy, not the exaggeration of discriminability.

2.2. Evaluations of Color-to-Grayscale Conversions

Apart from simple evaluations of the proposed methods surveyed below, we are not aware of any subjective perceptual evaluation study of color-to-grayscale conversions.

Bala and Eschbach [BE04] performed a small preference experiment to evaluate the qualitative performance of their conversion. The authors used three input color images that were converted using their novel method and by the simple conversion that retains the luminance component. The grayscale results were presented as hardcopy prints to six observers. The subjects preferred the novel spatial conversion approach (16 positive decisions out of total $6 \times 3 = 18$ comparisons).

Rasche et al. [RGW05] performed an accuracy experiment (with reference images) to assess their color-to-grayscale conversion. Six color images converted by the standard mapping of luminance to gray and by Rasche's method were presented to a group of 17 observers. The results revealed that for one group of input images the performance of the evaluated conversions was comparable, while for the second group of images, Rasche's method outperformed the traditional conversion.

3. Perceptual Experiments

In this section we describe the specific details of perceptual experiments that we have conducted to evaluate tested color-to-grayscale image conversions. We utilized the psychophysical technique of paired comparisons [Dav88], namely the two-alternatives forced choice (2AFC) experiment paradigm. We performed two experiments: in the first experiment (for accuracy), the grayscale images were presented along with the original (reference) color image, and in the second experiment (for preference), the subjects saw two grayscale images without any reference.

3.1. Evaluated Color-to-Grayscale Conversions

In total, we evaluated seven color-to-grayscale conversions, summarized in Table 1. When available, we utilized the codes provided by the authors for a particular conversion, but otherwise we implemented the conversion personally. All the conversions were run using default (constant) parameter settings (please, refer to Table 1 for numerical values). We decided to use constant parameters over all the input images for several reasons: first, to ensure comparable conditions for all the conversion methods involved; second, to reduce the number of images that are presented to subjects; and lastly not to bias the results by choice (tweaking of parameters) of an experimenter or an author (as different people may have a different sense of what is the best grayscale image).

3.2. Input Images

One of the advantages of a good-quality color-to-grayscale conversion is to give compelling results over a wide range of input images. We used 24 input color images in our study, with various motifs, origins, gamuts, etc. (the collection of these images is shown in Table 4 on Page 9). The images depict plants (images 9, 13, 23), foliage (22), fruits & vegetables (1, 10), portraits (11, 16), various photos (3, 4, 14, 15, 19), paintings (6, 20), cartoons (5, 21), color testing images (2, 7, 8, 12, 17), and computational images (18, 24). All the images were rescaled to maximally span 390×390 pixels for presentation purposes (to fit on the screen with the reference image) and also for the computational demands of several conversions.

3.3. Experimental Design

The evaluated images were displayed on a characterized and calibrated monitor EIZO S1910, a 19-inch LCD display, in native resolution 1280×1024 pixels. Calibration was performed by X-Rite GretagMacbeth Eye-One Display 2 colorimeter to D65, 120 cd/m^2 , and colorimetrically characterized by measured ICC profiles. The experimental images were presented on a neutral gray background with a luminance of 18% of the white point. The experimentation room was neutrally painted, darkened (measured light level: 4 lux), and observers sat approximately 70 cm from the display. All testing was performed approximately in the same

time of day (before noon) to avoid fatigue or other factors. The total of 121 observers took part in our experiments. The observers were both male and female between the ages of 18 to 41, and all of them reported to have normal, or corrected-to-normal vision. Each subject was verbally introduced to the problem before the experiment, as described in the following section.

3.4. Experimental Procedure

The design of the experiments followed the 2AFC approach [Dav88]. Specifically, we utilized the software ‘Ranker’ which is available at ranker.sourceforge.net. Every grayscale image was compared with every other grayscale image (see Table 5 on Page 10), i.e. for each input color image, we have $n(n-1)/2 = (7 \times 6)/2 = 21$ comparisons, where $n = 7$ evaluated conversions. With 24 input color images, we would need $24 \times 21 = 504$ trials, which would be prohibitive for each subject. Therefore, we ran a pilot study to assess the reasonable amount of trials for one observer (and to verify the setup as well). The pilot study indicated that eight sets of grayscale images (21 comparisons in each), i.e. the 168 trials, is an acceptable quantity for one observer without experiencing exhaustion and loss of concentration. With eight randomly selected sets (balanced design), the whole experiment took approximately 20 minutes per observer. The sequence of images and the position of images on the display (left or right) were randomized. The type of the experiment (accuracy or preference) was also randomized, however for a given observer it remained constant.

Experiment with a reference (accuracy): every time, two grayscale images were displayed along with the color original in the middle. Observers were asked to select the one of the two grayscale images that was closer in appearance to the original color image, i.e. to select the image that better reproduced the original. More specifically, the instructions stated: “Your task is to select the grayscale image that better matches the colors of the original color image.”

Experiment without a reference (preference): every time, two grayscale images were displayed. Observers were instructed to select the grayscale image that they preferred. Specifically, the instructions stated: “Your task is to select the preferred grayscale image from the presented pair.” Generally, accuracy (with reference) experiments were slightly more time-demanding with comparison to preference (without reference) experiments, and took 20 to 30 minutes per observer.

4. Results and Discussion

A total of 121 observers completed 20328 observations (pair-wise comparisons). Based on a post-test questionnaire, the results of two observers were excluded as outliers because of color vision deficiencies. In the following, we

Source of Variation	SS	d.f.	MS	F	p
conversion	105.6	6	17.6	185.4	≈ 0
experiment	0	1	0	0	≈ 1
input image	0	23	0	0	≈ 1
conversion \times experiment	2.8	6	0.5	4.9	10^{-4}
conversion \times input image	260.1	138	1.9	19.9	≈ 0
experiment \times image	≈ -0	23	≈ -0	≈ -0	≈ 1
Residual	13.1	138	0.1		
Total	381.5	335			

Table 2: The results of multi-factorial ANOVA test (where SS denotes Sum of Squares, d.f. means Degrees of Freedom, MS denotes Mean Square, F is F value, and p is p-value for the null hypothesis [TF07]).

present the results based on the observations of 60 participants who performed the accuracy experiment and 59 subjects who took part in the preference experiment. For each trial, the grayscale image chosen by an observer was given a score of 1, the other a score of 0. The data were stored in a 7×7 frequency matrix for each observer, where the value in column i and row j represents the score of grayscale conversion i compared with conversion j . We used Thurstone’s Law of Comparative Judgments, Case V, to convert the data into interval z-score (standard score) scales [Thu27, Eng00].

As the z-scores calculated from the observation data using Thurstone’s law are normally distributed, we can utilize classical parametric statistics in the further analysis. To inquire the significance of the *input images*, the *experiments* (accuracy and preference), and the *conversions* (i.e. the factors) on the observation data, it is profitable to apply the multi-factorial analysis of variance (ANOVA) test [TF07]. Multi-factorial (n-way) ANOVA is able to consider all the factors at once. The results of the n-way ANOVA are summarized in an ANOVA table [MR99] (Table 2). The results show that the only significant *main effect* is the conversion (because the p-value is below the threshold of 0.05), which means that there are significant differences in the performances of the inquired conversions. Neither the experiment type, nor the input image can alone explain the variability in the data. However, two statistically significant *interaction effects* imply that the observed scores depend on the combination of the conversion and the input image, and (with the smallest probability, but still with a statistical significance) on the combination of the conversion and the type of the experiment. This result suggests that the performances of the conversions depend on input images and on experiment type, and it makes sense to show the results separately for each input image and for each experiment. Finally, we performed a multiple comparison test (Tukey’s honestly significant differences [HT87]) over all the subjective data. This test re-

Decolorize	Smith08	CIE Y	Color2Gray	Rasche05	Neumann07	Bala04
0.544	0.487	0.158	0.149	-0.203	-0.317	-0.819

Figure 2: Overall performances of the inquired conversions. Results of the multiple comparison across all input images in both experiments. The best result is the leftmost, any conversions that are underlined are considered perceptually similar.

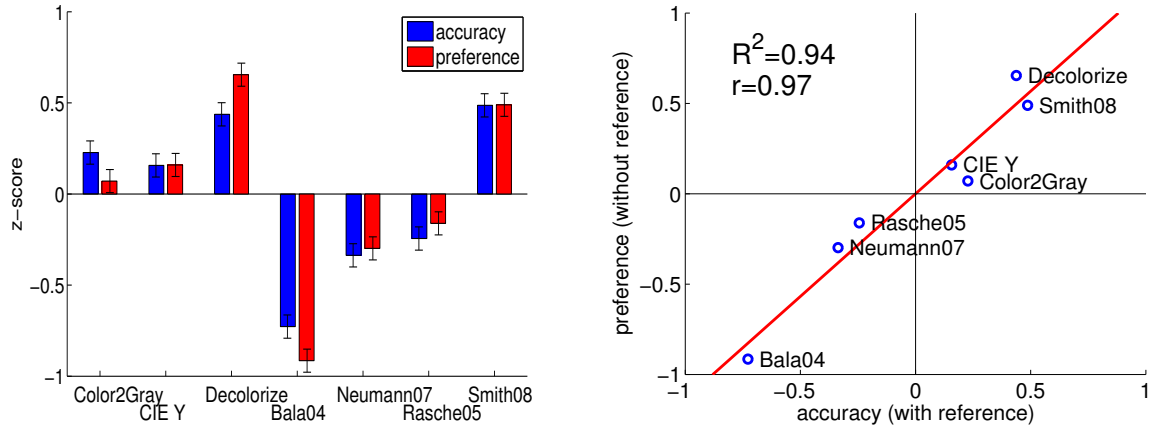


Figure 3: Overall results separately for the two experiments. Left: overall scores for both the accuracy and preference experiments. Error bars show intervals of 95% confidence. Right: comparison of accuracy and preference experiments.

turns an overall ranking of the conversions with the indication of the statistical significance of the differences between them (please, see Figure 2). The results show that the best ranked conversion in our study is Decolorize, but it performs statistically similar to Smith08; the worst ranked is Bala04.

4.1. Overall Accuracy and Preference Results

The overall scores were obtained by averaging the percentage matrices over all input images separately for the accuracy and preference experiments (please, see Figure 3). We can see from the overall results that altogether the best score in the accuracy experiment was achieved by Smith08, while Decolorize produces the most preferred grayscale images. Bala04 was ranked the worst in both the accuracy and preference experiments.

Comparing the overall accuracy and preference scores, we see similar trends in the results of the experiments. The calculated Pearson correlation coefficient [MR99] $r = 0.97$ and the coefficient of determination [MR99] $R^2 = 0.94$ (Figure 3 right) indicate high similarity of the preference and overall accuracy experiments. Notice that the CIE Y and Smith08 methods exhibit almost unchanged performance in both experiments. On the other hand, the rest of the methods show certain differences in accuracy and preference experiments. Specifically, Decolorize, Neumann07, and Rasche05 perform better in the preference experiment than in the accuracy experiment. On the contrary Color2Gray and Bala04 perform better in the accuracy experiment than in the preference experiment. Please refer to Section 4.3 for further analysis of accuracy and preference.

4.2. Results for Individual Images

Next, we examined the experimental data for all the color images individually (please see the summarized results in Table 3). We converted the observation data into z-scores independently for each input image using the Thurstone's Law of Comparative Judgments. The ranking reported in Table 3 is based on the calculated z-scores. The coefficient of agreement between subjects u ranges from $u = -1/(s-1)$, where s is the number of subjects, (which indicates no agreement between subjects) to $u = 1$ (all subjects responded the same). We show the results of the χ^2 test on the coefficient u , and the obtained p -values. The coefficient of consistency of subject's responses ζ ranges from $\zeta = 0$ (no consistency) to $\zeta = 1$ (ideally consistent responses), we report the average ζ over the subjects for a given input image. The values of u , ζ , χ^2 , and p were calculated in a similar way to Ledda et al. [LCTS05].

The results of the χ^2 test show that there is some agreement between observers, seen by the reported statistical significance (all the p -values of the null hypothesis are clearly below the threshold). This means that there are differences in performances of the conversions, which is also revealed by the ANOVA test reported above. The high values of ζ suggest that each subject was fairly consistent in their judgments. On the other hand, the agreement u amongst subjects varies from high values (images 2, 8) to lower agreement (for images 3, 9, 11), which indicates that the complexity of judgments differ depending on the input image.

Table 3 shows that no conversion produces universally satisfying results for all involved input images. Each of the

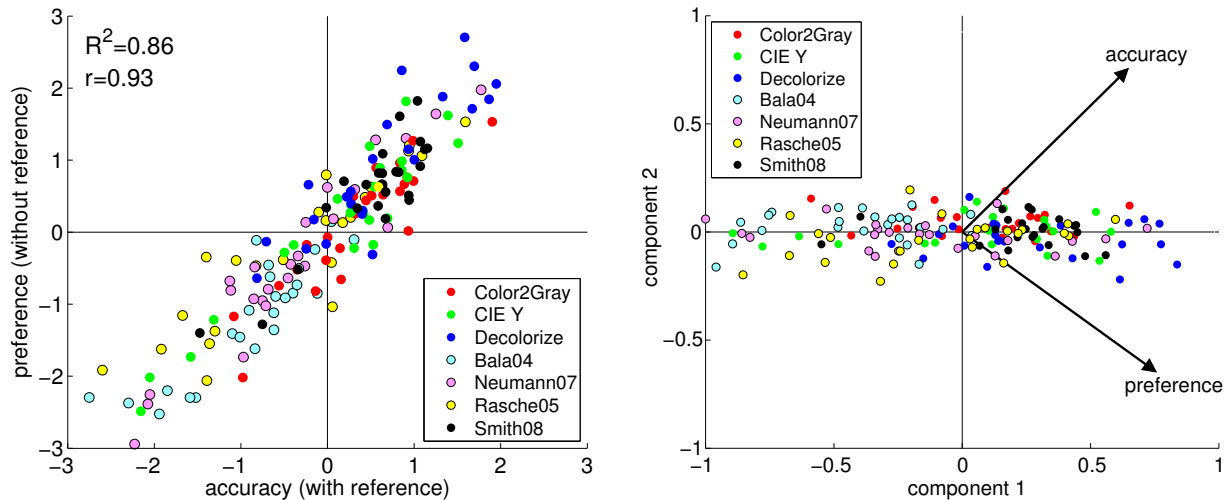


Figure 4: The results for individual input images. Left: Accuracy-preference scores for all input images. Each point represents a score of one color-to-gray conversion method for a particular input image. Right: Principal component analysis. The axes represent the principal components and the points represent the principal component scores of one conversion for one input image. The vectors show the values of principal component coefficients for the accuracy and preference variables.

seven tested conversions was ranked as worst for at least one input image and, apart from Bala04, each conversion was ranked as best for some input image. It is interesting to notice that Decolorize exhibits exceptionally good results for those input images that have rather narrow color gamuts or a limited number of colors (i.e. the images 2, 5, 7, 8, 10, 17, 20), refer to Table 4. For such images it is possible that, the image-dependent global mapping of Decolorize performs very well. Contrarily, Smith08 excels at colorful images with extensive color gamuts (4, 9, 15, 19, 22, 23), where the locally enhanced mapping based on the Helmholtz-Kohlrausch effect outperforms other conversions. Of note, however, is that the simple CIE Y conversion also performs quite well for these input images and it is remarkably good in images 5, 16, 18, 19, and 21.

4.3. Accuracy vs. Preference

We calculated values of the correlation coefficient r and the coefficient of determination R^2 [MR99] to determine the relationship between the accuracy and preference scores. The high values of r and R^2 for overall accuracy and preference scores (Figure 3, right) as well as for the scores for individual images (Figure 4, left) imply that there is a strong correlation between peoples' judgments of the color-to-grayscale conversion accuracy and the grayscale image preference. This suggests that one aspect dominates subjective judgment – let us call it an overall perceptual quality of color-to-grayscale conversion. The high values of correlations are interesting, as one would expect tricky judgments for grayscale pairs without the reference of some input images (e.g. 6, 7, 12, 17). The values of u and ζ , however, imply that the subjects were rather consistent in their opinions.

The principal component analysis [TF07] results in two principal components, where the first principal component explains 96.4% of the data variance (Figure 4, right). As illustrated, the first component (perhaps the overall quality of the conversion) lies nearly perfectly in the axis of accuracy and preference vectors. This result supports the above idea that only one dimension prevails in our subjective data.

4.4. Comparison to Previous Work

We believe that the presented study is much more credible than the two simple evaluations described in Section 2.2, as the number of subjects, input images and evaluated conversions is much higher. However, it is interesting and fair to compare the results obtained with the results of the previous evaluations.

In the preference experiment of Bala and Eschbach [BE04], Bala04 performed better than the mapping retaining the luminance. The authors used three input images (two of them are very similar to this study's image14 and image21). In our preference experiment, CIE Y and Bala04 performed similarly for image14, and Bala04 performed worse than CIE Y for image21. In our overall results, Bala04 performed worse than CIE Y, which is not consistent with findings of Bala and Eschnach. Besides the higher number of observers in our experiment, the discrepancy in the two studies is perhaps due to the different experimental setups, since Bala and Eschnach presented hardcopy prints and we utilized an LCD monitor.

Rasche's [RGW05] results show that for four input images, the performance of Rasche05 is comparable to the standard mapping of luminance. For another three images,

	u	ζ (avg)	χ^2	p (21 d.f.)	best	ranking of scores						worst
image1 (accuracy)	0.191	0.833	101.2	p<0.001	C	Y	S	B	D	N	R	
image1 (preference)	0.290	0.832	136.8	p<0.001	C	Y	S	D	R	B	N	
image2 (accuracy)	0.713	0.966	290.4	p<0.001	N	D	R	C	S	Y	B	
image2 (preference)	0.804	0.981	324.9	p<0.001	D	N	R	C	S	Y	B	
image3 (accuracy)	0.103	0.673	64.2	p<0.001	R	Y	B	S	N	C	D	
image3 (preference)	0.134	0.696	74.4	p<0.001	S	R	Y	B	N	C	D	
image4 (accuracy)	0.326	0.827	158.0	p<0.001	S	Y	N	D	C	B	R	
image4 (preference)	0.585	0.893	254.4	p<0.001	S	Y	N	D	C	B	R	
image5 (accuracy)	0.489	0.929	226.5	p<0.001	Y	D	S	C	B	R	N	
image5 (preference)	0.561	0.946	245.0	p<0.001	D	S	Y	R	C	B	N	
image6 (accuracy)	0.468	0.876	197.7	p<0.001	R	D	N	C	S	Y	B	
image6 (preference)	0.550	0.891	228.9	p<0.001	R	N	D	C	S	Y	B	
image7 (accuracy)	0.258	0.876	118.6	p<0.001	D	Y	R	C	B	N	S	
image7 (preference)	0.425	0.925	199.5	p<0.001	D	R	Y	C	B	N	S	
image8 (accuracy)	0.567	0.929	235.2	p<0.001	D	N	R	C	S	B	Y	
image8 (preference)	0.667	0.977	273.1	p<0.001	D	N	R	S	C	B	Y	
image9 (accuracy)	0.106	0.771	60.9	p<0.001	S	D	R	Y	B	N	C	
image9 (preference)	0.199	0.737	96.3	p<0.001	S	D	Y	R	N	B	C	
image10 (accuracy)	0.162	0.853	82.4	p<0.001	D	S	R	Y	N	B	C	
image10 (preference)	0.484	0.861	204.1	p<0.001	D	S	Y	R	N	B	C	
image11 (accuracy)	0.138	0.703	73.1	p<0.001	S	R	Y	D	C	N	B	
image11 (preference)	0.186	0.737	91.2	p<0.001	S	R	D	B	Y	N	C	
image12 (accuracy)	0.564	0.940	234.4	p<0.001	C	N	D	S	R	B	Y	
image12 (preference)	0.552	0.956	252.8	p<0.001	C	D	N	S	R	Y	B	
image13 (accuracy)	0.307	0.846	137.1	p<0.001	N	Y	S	C	D	B	R	
image13 (preference)	0.146	0.803	82.1	p<0.001	D	C	Y	S	N	B	R	
image14 (accuracy)	0.288	0.756	129.9	p<0.001	S	Y	C	D	N	B	R	
image14 (preference)	0.173	0.671	90.2	p<0.001	D	C	S	N	Y	B	R	
image15 (accuracy)	0.256	0.801	117.7	p<0.001	S	R	D	C	Y	B	N	
image15 (preference)	0.247	0.786	124.8	p<0.001	S	R	Y	C	D	N	B	
image16 (accuracy)	0.217	0.827	112.2	p<0.001	C	S	Y	R	D	N	B	
image16 (preference)	0.372	0.868	169.4	p<0.001	Y	S	C	D	N	R	B	
image17 (accuracy)	0.333	0.908	161.0	p<0.001	D	S	R	N	Y	B	C	
image17 (preference)	0.391	0.929	177.2	p<0.001	D	S	N	Y	R	B	C	
image18 (accuracy)	0.231	0.762	118.0	p<0.001	Y	S	C	D	B	N	R	
image18 (preference)	0.247	0.736	119.6	p<0.001	Y	S	C	D	R	N	B	
image19 (accuracy)	0.273	0.842	124.1	p<0.001	Y	S	C	D	B	N	R	
image19 (preference)	0.409	0.867	192.6	p<0.001	S	Y	C	D	B	R	N	
image20 (accuracy)	0.520	0.861	217.5	p<0.001	D	C	S	N	Y	R	B	
image20 (preference)	0.530	0.895	243.7	p<0.001	D	S	C	N	Y	R	B	
image21 (accuracy)	0.462	0.951	195.6	p<0.001	Y	S	D	C	N	B	R	
image21 (preference)	0.538	0.977	224.3	p<0.001	Y	D	S	C	N	B	R	
image22 (accuracy)	0.484	0.861	204.1	p<0.001	S	C	R	D	Y	N	B	
image22 (preference)	0.491	0.880	206.6	p<0.001	S	R	C	Y	D	N	B	
image23 (accuracy)	0.406	0.840	191.5	p<0.001	S	C	Y	B	R	D	N	
image23 (preference)	0.390	0.832	176.8	p<0.001	S	Y	C	R	B	D	N	
image24 (accuracy)	0.303	0.797	135.4	p<0.001	S	C	Y	D	R	B	N	
image24 (preference)	0.296	0.837	145.4	p<0.001	D	Y	C	S	R	B	N	

Table 3: The results for individual input images. Used abbreviations: avg=average, d.f.=degrees of freedom, C=Color2Gray, Y=CIE Y, D=Decolorize, B=Bala04, N=Neumann07, R=Rasche05, S=Smith08, notice that the used colors are equivalent to the colors in Figure 4.

the second one from the evaluated adjustments of parameters of Rasche05 outperforms the traditional conversion. In our experiment, the overall accuracy score of Rasche05 is close to CIE Y, but it is worse than CIE Y with statistical significance. Rasche05 outperforms CIE Y only for 11 of 24 input images (i.e. for images 2, 3, 6, 8, 9, 10, 11, 12, 15, 17, 22). The reason why Rasche05 performs worse in our study than in the Rasche experiment is due to the fact that we applied Rasche05 with constant parameters (alike the other conversions, seen in Table 1). We admit that Rasche05 could be ranked better after a thorough parameter tuning for each image (and other conversions, too), however this was not the objective of our study (please, refer to the discussion in Section 3.1).

5. Conclusions and Future Work

We presented a perceptual evaluation of color-to-grayscale image conversions. In two experiments, a total number of 119 subjects assessed the accuracy and the preference of grayscale images produced by seven state-of-the-art conversion methods. The inputs of the evaluated conversions represented the set of 24 color images of varying characteristics, motifs, and acquisitions.

The results show that the Decolorize [GD05] and Smith04 [SLJT08] conversions are overall the best ranked approaches, and the approach of Bala04 [BE04] performed the worst. However, the analysis of individual images reveal that no conversion produces universally good results for all the involved input images. Specifically, each of the seven inquired conversions was ranked the worst for at least one input image and, apart from Bala04, each conversion was ranked the best for some input image. These results suggest that there still exist areas for improvement of current conversions, especially in the robustness over various inputs. Furthermore, we found a high degree of correspondence between the accuracy and preference scores. Specifically, the results indicate that one dimension prevails in the subjects' judgment of the quality of the grayscale results. We believe that this is of particular importance and it is necessary to conduct experimental subjective studies, such as the one presented, to validate and evaluate color-to-grayscale conversions properly in order to expose their strengths and weaknesses, and to attain a deeper understanding of the examined field.

The presented study does not reflect computational demands, implementation difficulties, and other factors, which can play an important role for practical use. Notice that our results are valid for images presented on a screen, and the tested conversions may perform differently for hardcopy printouts or other media. Moreover, the desirable properties of the color-to-grayscale conversion may sometimes depend on the chosen application. In future work, we plan to implement all the conversions in the same platform to assess their computational demands and their actual usefulness. We will

also research how to involve more input parameters of the conversions so as to explore the parameter space.

Acknowledgements

This work has been supported by the Ministry of Education, Youth and Sports of the Czech Republic (research programs MSM 6840770014 and LC-06008), and by the Aktion OE/CZ grant no. 48p11. Special gratitude to M. Kalouš who coded the 'Ranker', Z. Mikovec, I. Malý, and O. Poláček for help in carrying out the experiments, J. Křivánek and J. Bittner for valuable comments, and to all the participants in the experiments for time and patience.

References

- [AK06] ALSAM A., KOLAS O.: Grey colour sharpening. In *Proc. of 14th Color Imaging Conf.* (2006), IS&T & SID, pp. 263–267.
- [BE04] BALA R., ESCHBACH R.: Spatial color-to-grayscale transform preserving chrominance edge information. In *Color Imaging Conference* (2004), IS&T, pp. 82–86.
- [Dav88] DAVID H. A.: *The Method of Paired Comparisons*, 2nd ed. Oxford University Press, 1988.
- [dQB06] DE QUEIROZ R. L., BRAUN K. M.: Color to gray and back: color embedding into textured gray images. *IEEE Trans. on Image Processing* 15 (June 2006), 1464–1470.
- [Eng00] ENGELDRUM P. G.: *Psychometric scaling: a toolkit for imaging systems development*, 1st ed. Imcotek Press, 2000.
- [Fai05] FAIRCHILD M. D.: *Color Appearance Models*, 2nd ed. Wiley-IS&T, Chichester, UK, 2005.
- [GD05] GRUNDLAND M., DODGSON N. A.: *The Decolorize Algorithm for Contrast Enhancing, Color to Grayscale Conversion*. Tech. Rep. UCAM-CL-TR-649, University of Cambridge, 2005.
- [GOTG05] GOOCH A. A., OLSEN S. C., TUMBLIN J., GOOCH B.: Color2Gray: saliency-preserving color removal. *ACM Trans. Graph.* 24, 3 (2005), 634–639.
- [GW02] GONZALEZ R. C., WOODS R. E.: *Digital Image Processing*, 2nd ed. Prentice-Hall, 2002.
- [HT87] HOCHBERG Y., TAMHANE A. C.: *Multiple Comparison Procedures*, 1st ed. Wiley, 1987.
- [LCTS05] LEDDA P., CHALMERS A., TROSCIANKO T., SEETZEN H.: Evaluation of tone mapping operators using a high dynamic range display. *ACM Trans. Graph.* 24, 3 (2005), 640–648.
- [MMS06] MANTIUK R., MYSZKOWSKI K., SEIDEL H.-P.: A perceptual framework for contrast processing of high dynamic range images. *ACM Trans. Appl. Perc.* 3, 3 (2006), 286–308.
- [MR99] MONGOMERY D. C., RUNGER G. C.: *Applied Statistics and Probability for Engineers*, 2nd ed. John Wiley & Sons, 1999.
- [NČN07] NEUMANN L., ČADÍK M., NEMCSICS A.: An efficient perception-based adaptive color to gray transformation. In *Proc. of Computational Aesthetics 2007* (2007), Eurographics Association, pp. 73–80.
- [RGW05] RASCHE K., GEIST R., WESTALL J.: Re-coloring Images for Gamuts of Lower Dimension. *Computer Graphics Forum* 24, 3 (2005), 423–432.
- [SLJT08] SMITH K., LANDES P.-E., JÖELLE THOLLOT K. M.: Apparent greyscale: A simple and fast conversion to perceptually accurate images and video. *Computer Graphics Forum* 27, 3 (2008).
- [TF07] TABACHNICK B. G., FIDELL L. S.: *Using multivariate statistics*, 5th ed. Pearson Education, 2007.
- [Thu27] THURSTONE L. L.: A law of comparative judgement. *Psychological Review* 34 (1927), 273–286.


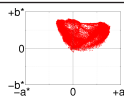

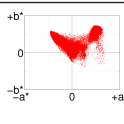
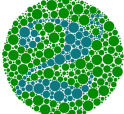
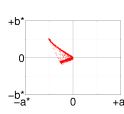

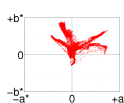

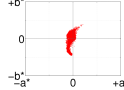

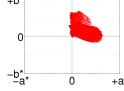

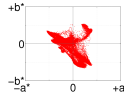

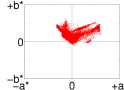

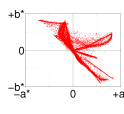



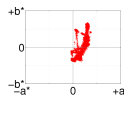
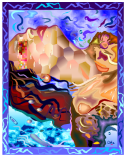
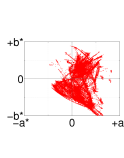

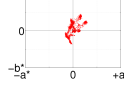

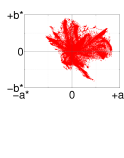

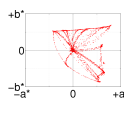

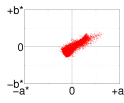

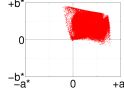
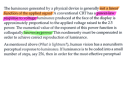
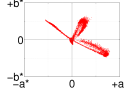

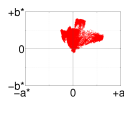

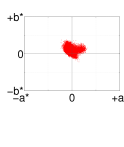

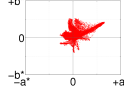

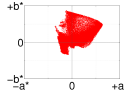
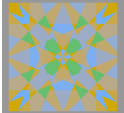
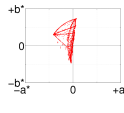

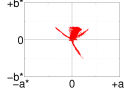
id	color image	color gamut	gamut [min, max]	id	color image	color gamut	gamut [min, max]
image1			$L^* = [0.06151, 100]$ $a^* = [-35.32, 77.37]$ $b^* = [-10.55, 81.26]$	image13			$L^* = [0.9717, 93.59]$ $a^* = [-46.35, 71.04]$ $b^* = [-38.17, 73.74]$
image2			$L^* = [38.67, 100]$ $a^* = [-51.06, 0.6223]$ $b^* = [-19.64, 52.48]$	image14			$L^* = [1.72, 98.19]$ $a^* = [-44.35, 74.49]$ $b^* = [-29.15, 90.95]$
image3			$L^* = [0, 100]$ $a^* = [-15.18, 20.87]$ $b^* = [-45.69, 38.37]$	image15			$L^* = [5.421, 99.72]$ $a^* = [-6.689, 64.22]$ $b^* = [-17.33, 69.62]$
image4			$L^* = [0.7095, 99.71]$ $a^* = [-54.93, 40.19]$ $b^* = [-66.35, 53.9]$	image16			$L^* = [0, 99.3]$ $a^* = [-32.98, 60.36]$ $b^* = [-17.39, 61.53]$
image5			$L^* = [0, 100]$ $a^* = [-71.94, 84.66]$ $b^* = [-92.34, 83.02]$	image17			$L^* = [56.07, 60.27]$ $a^* = [-1.697, 61.24]$ $b^* = [-38.39, 42.2]$
image6			$L^* = [14.67, 96.38]$ $a^* = [-5.309, 38.68]$ $b^* = [-41.72, 68.38]$	image18			$L^* = [0, 100]$ $a^* = [-46.81, 82.9]$ $b^* = [-112.1, 88.87]$
image7			$L^* = [64.75, 100]$ $a^* = [-16.63, 30.27]$ $b^* = [-36.28, 45.12]$	image19			$L^* = [0, 100]$ $a^* = [-55.68, 83.98]$ $b^* = [-81.79, 90.77]$
image8			$L^* = [42.24, 57.86]$ $a^* = [-42.78, 79.86]$ $b^* = [-87.36, 69.06]$	image20			$L^* = [8.564, 81.58]$ $a^* = [-26.9, 65.64]$ $b^* = [-33.65, 40.39]$
image9			$L^* = [0.4412, 100]$ $a^* = [-21.38, 79.57]$ $b^* = [-14.87, 91.16]$	image21			$L^* = [3.012, 100]$ $a^* = [-55.65, 78.98]$ $b^* = [-47.86, 64.1]$
image10			$L^* = [0, 100]$ $a^* = [-25.91, 64.11]$ $b^* = [-11.55, 81.48]$	image22			$L^* = [3.13, 100]$ $a^* = [-23.04, 31.68]$ $b^* = [-22.23, 37.5]$
image11			$L^* = [0, 100]$ $a^* = [-28.37, 66.11]$ $b^* = [-40.77, 52.23]$	image23			$L^* = [0.7857, 98.24]$ $a^* = [-35.58, 63.81]$ $b^* = [-34.31, 87.03]$
image12			$L^* = [62.76, 71.36]$ $a^* = [-41.11, 7.765]$ $b^* = [-46.83, 73.57]$	image24			$L^* = [0, 99.9]$ $a^* = [-34.36, 34.24]$ $b^* = [-50.49, 35.87]$

Table 4: The set of input images. Images courtesy of e-cobo.com (1), A. Gooch (2, 7, 8, 17), R. E. Barber (3), K. Rasche (4, 13, 22), imagekingdom.com (5), L. Neumann (6, 9, 12), Kodak (11, 14), UT Austin (15), Sony (16), Fujifilm (19), artcyclopedia.com (20), M. Čadík (21), and K. Odhner (24).

id	color image	CIE Y	Color2Gray	Decolorize	Smith08	Rasche05	Bala04	Neumann07
image1								
image2								
image3								
image4								
image5								
image6								
image7								
image8								
image9								
image10								
image11								
image12								
image13								
image14								

Table 5: The results of the evaluated color-to-grayscale conversion methods. Please, refer to the accompanying webpage: http://www.cgg.cvut.cz/~cadikm/color_to_gray_evaluation for the complete set of the full-resolution images.