Experimental Approach for Human Perception Based Image Quality Assessment

Jin-Seo Kim[1], Maeng-Sub Cho[1], and Bon-Ki Koo[1]

[1] CG Research Team, Digital Content Research Division,
Electronics and Telecommunications Research Institute,
161 Gajeong-dong, Yuseong-gu, Daejeon, 305-350, Republic of Korea
{kjseo, choms, bkkoo}@etri.re.kr
http://dcon.etri.re.kr

Abstract. The term 'image quality' is a subject notion so it is difficult to quantify. However, it can be reasonably quantified by using statistical and psychophysical approaches. Furthermore, it is also possible to model the human perception of image quality. In this paper, large scale psychophysical experiments including pair comparison and categorical judgment were carried out to judge the perception of image quality of photographic images. The evaluation of both image difference and absolute quality was also carried out. Test images were generated by rendering the eight selected original images according to the change of lightness, chroma, contrast, sharpness and noise attributes. Total number of 288 images were used as test images. The experimental results were used to calculate z-scores and colour difference threshold to verify the optimum level for each transform function. User preferred image content can be provided to entertainment, education, etc. when using the result of the study.

Keywords: Image quality, CIELAB, Psychophysical experiment

1 Introduction

Image quality is an ideal concept and, therefore it can be determined by many different attributes such as colour, resolution, sharpness and noise. A number of metrics have been published that could be used to predict image quality including CIECAM02 [1], iCAM [2], MTFA, SNR and MSE. However, none of these metrics can easily predict certain perceptual attributes of human vision such as the naturalness of the image [3]. CIE TC8-02 is studying the calculation of colour difference using spatial characteristics.

The aim of this study is to derive a colour-appearance model which can predict both the spatial and subjective attributes of image quality (sharpness, noise, naturalness, etc.) so that many of image content based applications can provide the best quality image content to their users. To determine image-quality attributes psychophysical experiments have been conducted and the performance of current colour-difference formulae was evaluated. Six attributes were evaluated in this study (lightness, chroma, contrast, noise, sharpness and compression) using CIELAB and S-CIELAB to calculate thresholds. CIELAB is one of the CIE standard colour spaces in which the Euclidean distance between two points in the CIELAB space is considered as colour difference. S-CIELAB is the updated version of CIELAB so that the spatial attributes can be considered when calculating the colour difference of two images, original and spatially corrupted image.

2 Experimental Method

Psychophysical experiments were conducted in order to collect the individual preference data of some test images for the development of image-quality modelling algorithm. A BARCO Reference Calibrator®121 was used in a darkened room as a reference display device for the experiments. Some device characteristics such as spatial and temporal uniformity and the channel additivity were tested and found to be satisfactory for conducting a psychophysical experiment. The GOG model was used to characterise the display used in the experiment [4].

Two types of psychophysical experiments were carried out; pair comparison and categorical judgment. Pair comparison was conducted for the evaluation of appearance difference between pairs of sample images. Categorical judgment which uses single image was also conducted for the evaluation of naturalness of individual test images.

Eight different test images were chosen to represent photo-realistic images (e.g. fruit, foliage, flower, plant) and artificial objects (e.g. balloon, bicycle, clothes). Fig. 1 shows the test images used in the experiment.

(a) Musician (b) Fruit (c) Metal wares (d) Balloons

(e) Bicycle (f) Happy girl (g) Mirror image (d) Chair

Fig. 1. Test images.

Six image-quality attributes (lightness, chroma, contrast, noise, sharpness, and compression) were chosen in this study and, six different levels of transform for each attribute were applied to prepare test images. Total numbers of 36 rendered images were generated as a result. The colour transform functions used in the experiments are summarised in Table 1.

Table 1. Image quality transformation functions.

Parameter	**Lightness**	**Chroma**
Formula	$L^*_{out} = kL^*_{in}$ *k:* scaling factor	$C^*_{out} = kC^*_{in}$ *K:* scaling factor
Abb.	L	C

Parameter	**Contrast**	**Noise**
Formula	$L^*_{out} = L^*_{mid} + L^*_{in}\times k$, *where,* $L^*_{in} \geq L^*_{mid}$ $= L^*_{mid} - L^*_{in}\times k$, *where,* $L^*_{in} < L^*_{mid}$ $C^*_{out} = C^*_{mid} + C^*_{in}\times k$, *where,* $C^*_{in} \geq C^*_{mid}$ $= C^*_{mid} - C^*_{in}\times k$, *where,* $C^*_{in} < C^*_{mid}$ *k:* scaling factor L^*_{mid}: average lightness of the image C^*_{mid}: average chroma of the image	Gaussian random noise
Abb.	CLC	N

Parameter	**Sharpness**	**Compression**
Formula	3×3 mask	Adobe photoshop's jpeg compression function
Abb.	SB	CO

Example of transformed images for lightness rendering is shown in Fig. 2

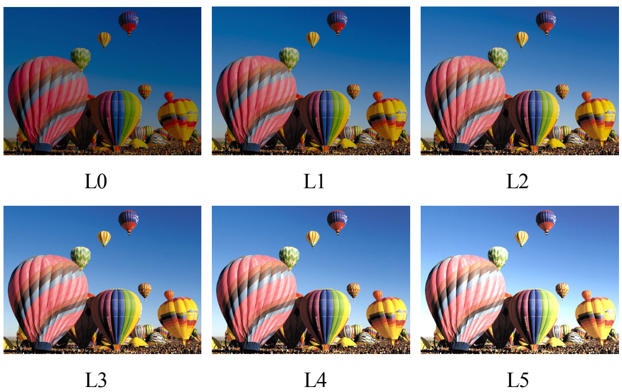


Fig. 2. Six different lightness transformed images.

Total number of 288 rendered images (8 images × 6 parameters × 6 levels) plus eight original images were prepared as test images. Overall, 18 observers participated in the pair-comparison experiment and 11 observers participated in the categorical-judgment experiment. All observers were tested and found to have normal colour vision. For the pair-comparison experiment, the original and one of the transformed images were displayed on a CRT, and observers were asked the questions listed in Table 2. A total number of 2,304 observations (8 images × 6 parameters × 6 levels × 4 questions × 2 repeats) were obtained for each observer. For the categorical-judgment experiment, a single image (either the original or one of the transformed ones) was displayed on a CRT in a random order, and observers were asked to assign a number from a scale 1-9 for equally stepped categories according to the questions listed in Table 2.

Table 2. Questions used in the experiments.

Experiment	Pair comparison	Categorical judgment
Questions	1. Do they look the same? (overall) 2. Do they look the same in colour? 3. Do they look the same in sharpness? 4. Do they look the same in texture?	1. How real is this image? (overall) 2. How real is the colour of this image? 3. How real is the texture of this image?

The experiments were divided into four sessions so that the observation time for any one session did not exceed 45 minutes in order to avoid fatigue. In total, 63,648 observations (41,472 for pair comparison and 22,176 for categorical judgment) were accumulated over one month.

The software tool was developed to carry out psychophysical experiments. It consisted of three parts; user information input, pair comparison experiments, and categorical judgment experiments. First, each observer should complete the user information part before starting the experiment. Then one of the two psychophysical experiments was carried out according to the pre-designed schedule. In Fig. 3, actual images for the experiments using the software were shown.

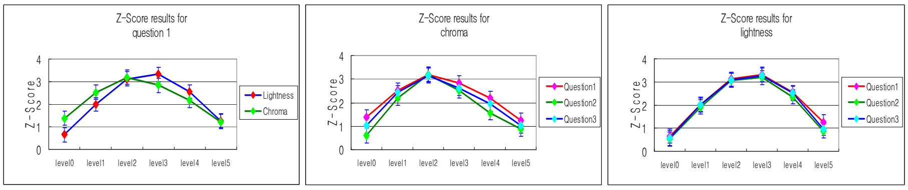
(a) User input dialogbox (b) Pair comparison (c) Categorical judgment

Fig. 3. Software tool for the experiment.

3 Data Analysis

Two types of experiments were carried out and the results of the data analysis are summarised below.

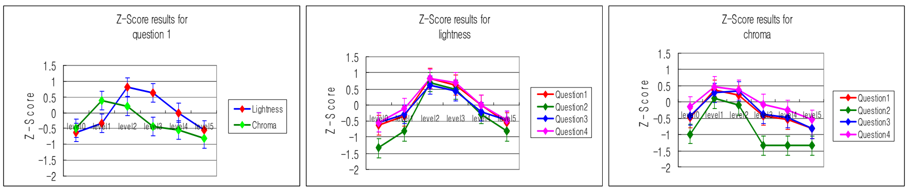
For the categorical-judgment experiment, z-scores were calculated to evaluate the image quality of different levels of colour-transformed images. Fig. 4 shows the z-score results for categorical judgment of the 'balloons' image. Fig. 4(a) is the z-score results of lightness- and chroma-transformed images to the question 1; How real is the image? Fig. 4(b) is the chroma-transformed results for all three questions. Fig. 4(c) is the lightness transformed results for the same three questions. It can be seen from Fig. 4 that the results for lightness and chroma show a similar characteristics, i.e. the highest image quality occurs in the middle of lightness or chroma levels (that is, for images close to the original). This suggests that photographed images with small colour transformations applied tend to match best with the memory colour, so that they might have the highest image quality scores.



(a) (b) (c)

Fig. 4. Z-score results for balloons test images – Categorical judgment.

Z-score analysis was done for pair comparison experiment. Fig. 5 shows the results.



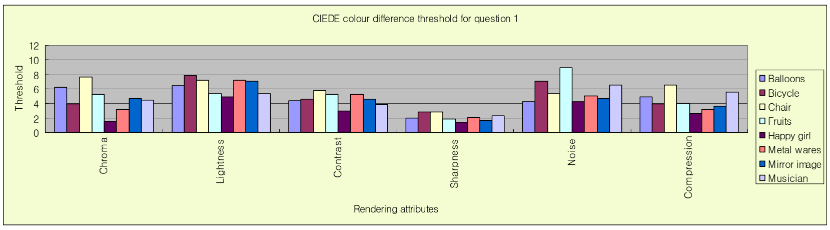
(a) (b) (c)

Fig. 5. Z-score results for balloons test images – Pair comparison.

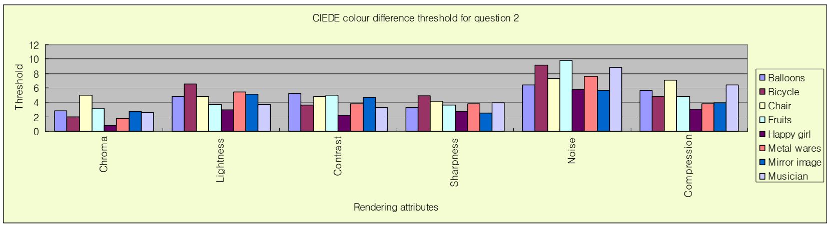
In Fig. 4, the highest score for chroma rendering is level 2 which is slightly less chromatic than the original whereas the highest scores for other rendering attributes except for noise and compression rendering are level 3 which is slightly emphasized in each rendering attribute. In all of the 7 test images except fruit image, similar results were obtained. In fruit image, level 3 has the highest scores in chroma rendering. On the contrary, highest scores are distributed in level 2 and level 3 randomly in pair comparison showed in Fig. 5 except chroma rendering which has the highest score in level

2, and this is the same result as categorical judgment. In all of the 7 test images except fruit image, similar results were obtained. This means human perception of image quality is image dependent when the reference image is shown simultaneously with the test image. From these z-score analysis, it can be assumed that, observers recognise images as optimum quality when the image attributes exhibit slightly more than the original in case of determining the image quality with single test image except for the chroma attribute which observers recognise high image quality when the attribute exhibits slightly less than the original. However, when the original images are shown with the rendered images, this phenomenon disappears and the rendered images with attributes in either slightly more or less than the original images are selected as the highest image quality. The reason can be thought from the experimental results for pair comparison showed in Fig. 5 that observers pay more attention to discriminating the textural difference rather than colour and other attributes when both original and test images are displayed simultaneously. Also texture of the image has higher correlation with the overall image quality than other attributes. Details are explained in next data analysis. Besides, in case of fruit image, people recognise high image quality when test images have little more chromatic attribute than the original. That means people have memory colours about fruits which have more chromatic than the original. In noise and compression rendering, level 0 has the smallest attribute change, so the plots are different from other attributes.

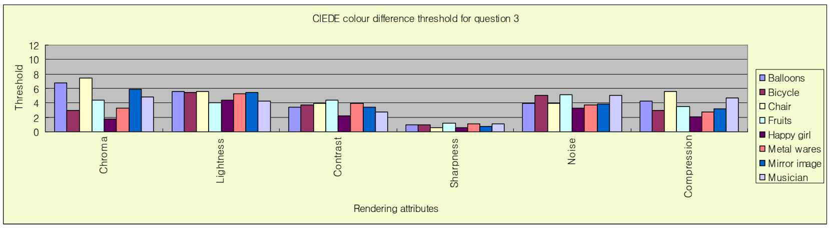
For pair-comparison experiment, colour-difference thresholds for each rendering attribute were calculated based upon CIELAB and S-CIELAB colour differences and Fig. 6, and Fig. 7 respectively. In Fig. 7, it was found that the 'Mirror' image had the highest threshold for most of the questions and the 'Happy girl' image had the lowest threshold for all questions. This means that people are less sensitive to the colour change in the 'mirror' image which includes natural objects such as tree, green foliage, blue sky, etc., while people are more sensitive to the 'happy girl' image which includes skin tones. In addition, the lightness has in general higher threshold values than the chroma thresholds. This implies that chroma differences are more noticeable than lightness differences, in agreement with earlier findings by Sano *et al*. [5], [6], [7]



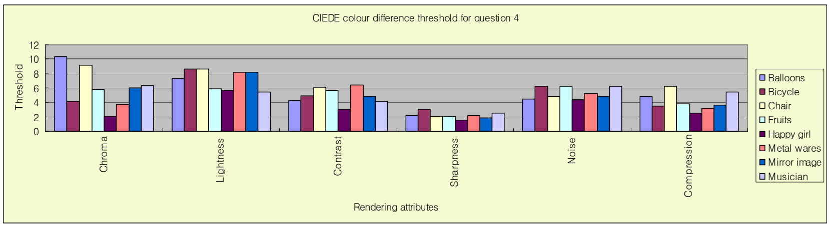
(a) Image quality difference



(b) Colour diference

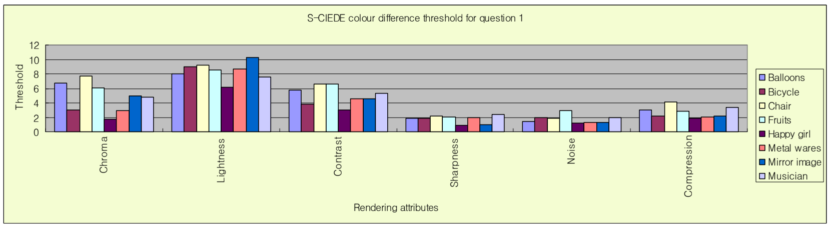


(c) Sharpness difference

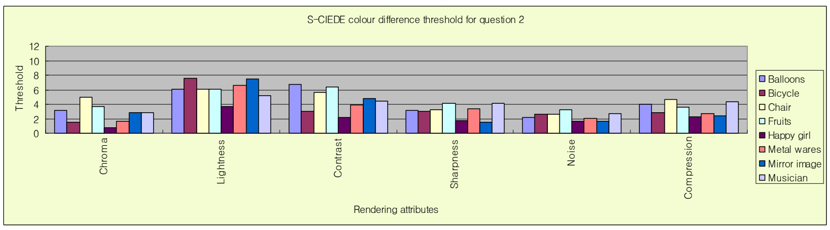


(d) Texture difference

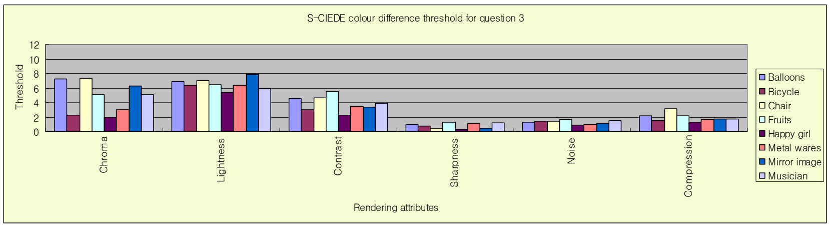
Fig. 6. Colour difference thresholds for each question - CIELAB.



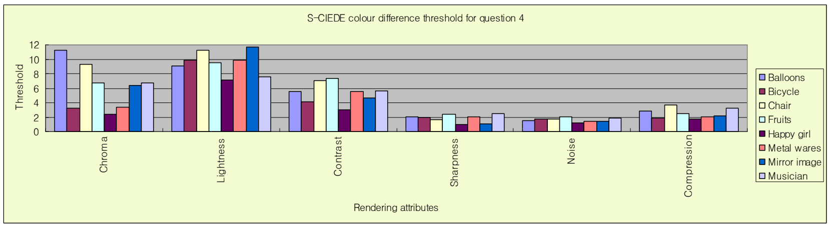
(a) Overall quality difference



(b) Colour difference



(c) Sharpness difference



(d) Texture difference

Fig. 7. Colour difference thresholds for each question – S-CIELAB.

Furthermore, S-CIELAB has higher thresholds for lightness and chroma rendering, whereas CIELAB has higher thresholds for sharpness, noise and compression rendering as shown in Fig. 6. And bicycle, chair and mirror image have relatively high threshold for lightness and chroma rendering in both CIELAB and S-CIELAB formulae while happy girl has lowest or relatively lower threshold for the remaining rendering in both CIELAB and S-CIELAB. This means that people have low sensitivity in discriminating the change of attributes the artificial objects such as bicycle, chair and high sensitivity in skin tone. The reason can be thought as people have higher sensitivity in recognising the change of human skin, and people have memory colours with slightly more chromatic of the objects than the real objects. [8], [9], [10]

Finally, the coefficient of variation (CV) defined as CV = (standard deviation / mean value) × 100 was calculated for each colour-difference formula in order to determine the performance of the formulae, CIELAB and S-CIELAB. Table 3 shows the results of CV calculation between CIEDE and SCIEDE for six different rendering functions. For a perfect agreement between the formula and visual results, CV should be zero.

Table 3. CV results for difference questions.

	CIEDE chroma	**SCIEDE chroma**	**CIEDE lightness**	**SCIEDE lightness**
Overall	41	44	17	15
Colour	48	50	25	21
Sharptness	42	45	13	11
Texture	47	50	19	17

	CIEDE contrast	SCIEDE contrast	CIEDE sharpness	SCIEDE sharpness
Overall	20	26	25	31
Colour	26	34	21	32
Sharpness	21	27	29	45
Texture	23	26	21	32

	CIEDE noise	SCIEDE noise	CIEDE compression	SCIEDE compression
Overall	29	32	30	30
Colour	21	24	28	28
Sharpness	17	21	31	30
Texture	15	17	30	29

It can be seen from Table 3 that the two formulae tested gave very similar result, although the CIELAB formula had a slightly better performance in predicting colour difference whereas the S-CIELAB formula was slightly better at predicting lightness and compression changes. The reason is that S-CIELAB uses spatial filtering that can predict spatial attributes such as compression, so its performance is better than CIELAB which only deals with colorimetric attributes of the pixel values. It was expected that most formulae would give similar performance because the transformed images used in the experiment had systematic spatial variations. The results also imply that all images had more or less the similar lightness threshold but large variations in other thresholds. In other words, people are less sensitive to lightness change than other attribute changes such as chroma, contrast, and sharpness changes.

4 Conclusion

An experiment was carried out to evaluate the image quality of colour-transformed images to test the performance of the CIELAB and S-CIELAB colour-difference formulae. Eight selected images were used and six colour-transform functions were generated. Each function had 6 distinct levels for rendering the images. Z-scores and colour difference thresholds were calculated from the original and 288 rendered images. The results reported here only include the data analysis of two colour-transform functions, lightness and chroma. The conclusions are summarised below and subsequent data analysis will be carried out for the remaining functions and reported elsewhere.

1) The results for the categorical judgment were similar for all the three questions asked.
2) People prefer a slightly lighter and higher chroma image to a darker and lower chroma one.
3) All images had similar lightness thresholds but large variations in chroma

thresholds.

4) The performances of CIELAB and S-CIELAB were similar, but lightness and chroma attributes for each formula have different CV results.

Based on this study, future study will cover the development of image quality prediction model and apply it to the digital image applications such as computer game, digital cinema, digital broadcasting so as to provide the user preferred image content. Also more image quality factors which may affect the image quality of moving images such as temporal frequency will be considered as an advanced research in the future.

References

1. Moroney N., Fairchild M.D., Hunt R.W.G., Li C., Luo M.R., Newman T.: The CIECAM02 Colour Appearance Model, Proceedings of the tenth Colour Imaging Conference. IS&T/SID, Scottsdale, Arizona, (2002) 23-27.
2. Fairchild M.D, Johnson G.: Meet iCAM: A next-generation colour appearance model, Proceedings of the tenth Colour Imaging Conference, IS&T/SID, Scottsdale, Arizona, (2002) 33-38.
3. Yendrikhovskij S.: Towards perceptually optimal colour reproduction of natural scenes, Colour Imaging Vision and Technology (Wiley, 1999), Chapter 18.
4. Berns S.: Methods for characterizing CRT displays, Displays vol. 6, no.4, (1996) 173-182
5. Sano C., Song T., Luo M.R.: Colour Differences for Complex Images, Proceedings of the eleventh Colour Imaging Conference, IS&T/SID, Scottsdale, Arizona, (2003) 121-125.
6. Uroz J., Luo M.R., Morovic J.: Perception of colour difference between printed images, Colour Science: Exploiting digital media, John Wiley & Sons Ltd., (2002) 49-73
7. Song T., Luo M.R.: Testing colour difference formulae on complex images using a CRT monitor, IS&T SID 8th Colour Imaging Conference, (2000) 44-48
8. Coren S., Ward L.M., Enns J.T.: Sensation and perception, Six edition, Wiley, pp. 114-115 (2004)
9. Wichmann, F. A., Sharpe, L. T., Gegenfurtner, K. R.; Contributions of colour to recognition memory for natural scenes. Journal of Experimental Psychology: Learning, Memory & Cognition, 28, (2002) 509-520
10. Newhall, S. M., Burnham, R. W., Clark, J. R.: Comparison of successive with simultaneous colour matching. Journal of the Optical Society of America, 47, (1957) 43-56 (1957)