An Evaluation of Colour-to-Greyscale Image Conversion by Linear Anisotropic Diffusion and Manual Colour Grading

Aldo Barba, Ivar Farup, Marius Pedersen; Department of Computer Science, Norwegian University of Science and Technology (NTNU), Gjøvik, Norway

Abstract

In the paper "Colour-to-Greyscale Image Conversion by Linear Anisotropic Diffusion of Perceptual Colour Metrics", Farup et al. presented an algorithm to convert colour images to greyscale. The algorithm produces greyscale reproductions that preserve detail derived from local colour differences in the original colour image. Such detail is extracted by using linear anisotropic diffusion to build a greyscale reproduction from a gradient of the original image that is in turn calculated using Riemannised colour metrics. The purpose of the current paper is to re-evaluate one of the psychometric experiments for these two methods (CIELAB L^* and anisotropic ΔE_{99}) by using a flipping method to compare their resulting images instead of the side by side method used in the original evaluation. In addition to testing the two selected algorithms, a third greyscale reproduction was manually created (colour graded) using a colour correction software commonly used to process motion pictures. Results of the psychometric experiment found that when comparing images using the flipping method, there was a statistically significant difference between the anisotropic ΔE_{99} and CIELAB L^* conversions that favored the anisotropic method. The comparison between ΔE_{99} conversion and the manually colour graded image also showed a statistically significant difference between them, in this case favoring the colour graded version.

Introduction

A common process in image processing and reproduction is the conversion of a colour image to greyscale. Both global and local approaches (or a combination) are normally used to perform such conversions. Since the process itself involves reducing the number of channels in the image, loss of information and image detail are common side effects.

To improve the quality of the resulting greyscale images, many methods have been researched. The most common method is to use the lightness channel (L^*) of the image represented in the CIELAB colour space. Being a global method, it involves generating one lightness channel that is a weighted sum of the RGB channels that ignores local variations in the image [1].

Another approach is to use what is called a spatial method that takes into account the variation within local areas of the image. We can understand this variation as a gradient (or rate of change) and measure it. In our case this is done using a Riemannian metric obtained from a colour difference formula calculated in a curved space [2] that is better suited to measure such perceptually uniform distances [3].

In the original paper by Farup et al. [1], upon which this work is based, several metrics were converted to this Riemannian space and then used to create a tensor for the colour image from its colour difference values. Eigenvalues and vectors from this colour image tensor were then used to construct a gradient, and from this gradient the final greyscale reproduction is constructed using both isotropic and linear anisotropic diffusion.

In addition to the conversions obtained using several colour metrics (ΔE_{ab}, ΔE_{uv}, ΔE_{00}, ΔE_{E}, ΔE_{99} (4 versions), and hyperbolic $\Delta E_{99\mathrm{c,hyp}}$) the proposed algorithm was also compared with the greyscale conversion methods proposed by Smith et al. [4] and Du et al. [5]. More details are found in the "Background" section or in Farup et al. [1].

In this re-evaluation we focus on the performance of the algorithm in the preference psychometric experiment done as part of the original paper. In that experiment, none of the methods tested performed significantly better than the CIELAB L^* conversion method. Additionally, given the performance of L∗ and ΔE99 in the original evaluation, and to take advantage of the experience of one of the author of the present paper as a colourist, it was decided to test both conversions against another global method: a manual greyscale reproduction.

This paper is organized as follows: first we present relevant background. Then we introduce the methods used. Results and discussion are presented before we conclude and propose future work.

Background

As part of the larger study carried out by Farup et al. [1], a preliminary study was included where the proposed algorithm was run on 30 images with two gradient creation alternatives, using both isotropic and anisotropic diffusion. In total nine different colour difference formulas were tested: Δ*E*ab, Δ*E*uv, Δ*E*00, Δ*E*E, Δ*E*99 (4 versions), and hyperbolic Δ*E*99c,hyp. As a result from this evaluation, the following was decided for the main experiment by Farup et al. [1]:

- Only one gradient creation method was selected.
- Anisotropic diffusion was selected over its isotropic counterpart, because of its better performance. Anisotropic diffusion was without the halo artefacts usually produced by
- isotropic diffusion.
- Euclidean Δ*E*99c was chosen over its hyperbolic alternative
- Δ*E*99c,hyp, because no visible or measurable difference between them could be found and therefore the original version was selected.
- Euclidean Δ*E*99 [6] was the found to be the more accurate metric to use according to a small psychometric experiment with nine images designed to select the best parameters for the proposed algorithm.

Following the preliminary study, a test data set was created with the greyscale reproductions from five algorithms chosen: ΔE_{99} (overall more accurate), ΔE_{00} (regarded as the best colour difference metric), CIELAB L^* (the most commonly used method), the method proposed by Smith et. al. (evaluated as most accurate in their study)

[4], and the method proposed by Du et. al. (also evaluated as superior in another psychometric experiment) [5]. These five algorithms were applied on 10 images: five from the CSIQ [7] data set and five from the Kodak data set. Two psychometric evaluations were performed on this test data set using the QuickEval [8] platform:

- Accuracy experiment: a paired comparison under con- trolled conditions in which observers were asked to "select the most accurate reproduction of the colour image" [1].
- Preference experiment: an online uncontrolled paired comparison where observers were asked to "select the image you prefer" [1].

It was found that linear anisotropic diffusion performed better than or equal to all the tested colour-to-greyscale conversion algorithms both in terms of preference and accuracy. Surprisingly, no significant difference was found between the sophisticated algorithms and a simple $L*$ luminance map.

Methods

In this paper we re-evaluate the preference experiment un- der a controlled environment (as opposed to an online experiment) using only the CIELAB L^* and anisotropic Euclidean ΔE_{99} reproductions for the pair comparison. Additionally, using the same setup and methodology we will evaluate both CIELAB L^* and ΔE_{99} against our manually created reproduction.

Colour-to-Grey Algorithms

Main Colour-to-Greyscale Algorithm

Since the scope of this paper did not involve the mathematical analysis or computational implementation of the algorithm itself, we refer to the original paper [1] for the details on the concepts dealt with when using linear anisotropic diffusion together with Riemannised colour metric differences for the creation of the greyscale reproductions.

Manual Colour-to-Greyscale Reproduction

Given the opportunity to conduct a controlled psychometric experiment and the previous experience of one of the authors of the present paper as a film colourist, it was agreed to test a third method against the lightness channel of CIELAB and the ΔE_{99} conversion.

The images in the data set were manually converted to greyscale using Blackmagic Design Davinci Resolve colour correction software. The conversion was approached as a global operation (not optimized for local detail preservation) by using a tool provided by the software called channel mixer that allows for the combination of weighted RGB channels into a final greyscale image.

The main goal during the creation of this third reproduction was to obtain a pleasing image in which the colours are represented by greyscale values perceived as their natural fit. For ex- ample skin tone was mapped to 50% lightness approximately, as such is the value for skin that is commonly used in photography. This approach is somewhat similar to CIELAB, as long as the image has been correctly exposed according to standard photo- graphic conventions.

Psychometric Preference Experiments

Following the setup as used in the paper by Farup et al. [1], the re-evaluation experiment took place in a dark surround, using an Eizo ColorEdge CG246W monitor which was hardware calibrated to sRGB (80 cd/m2 and 6500 K) using an xRite i1 pro spectrophotometer. The viewing distance of the experiment was set to approximately 50 cm. All observers had normal colour vision and normal or corrected to normal visual acuity.

Images were shown one at a time, giving the user the control to flip between the pair of greyscale reproductions being com- pared (as shown on the right side of Figure 1). This is the main difference with the original experiment, which instead presented the reproductions side by side (as shown on the left side of Figure 1). Since the main benefit of the proposed algorithm is to preserve detail from the colour difference information in the original images, this setup allowed for an easier way to spot the differences between reproductions for the observer (see Figure 1). The experimental setup is similar to that used in [9, 10], where observers could flip between the images.

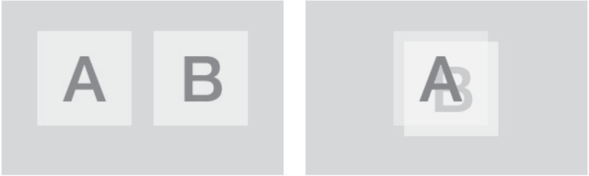

Figure 1. Experimental setup. On the left a traditional pair comparison setup with two reproductions (A and B) next to each other, which was used in the work by Farup et al. [1]. On the right the setup used in this work, where the images are placed on "top" of each other at the same location and the observers could flip between them.

The same 10 images (Figure 2) as used by Farup et al. [1] was included in the experiment. Five of these images are from the CSIQ dataset [7] and five images from the Kodak data set. The images were chosen following the recommendations from Field [11] in order to ensure a variety of image characteristics. For an overview of all the reproductions used see Figures 6 and 7.

CIELAB L vs. Anisotropic Diffusion of ∆E99

The experiment was done by 15 observers: 7 male and 8 female. In this group there were 9 colour science students, 3 interaction design students, 2 post doctoral researchers and 1 staff professor of computer science.

CIELAB L∗ vs. Manual Reproduction

This experiment was done by 14 observers: 6 male and 8 female. This group included 8 colour science students, 3 inter- action design students, 2 post doctoral researchers and 1 staff professor of computer science.

Anisotropic Diffusion of ∆E99 vs. Manual Reproduction

This experiment was done by 11 observers: 10 male and 1 female. This group included 1 colour science student, 8 post doctoral researchers, and 2 staff professors of computer science.

Data Analysis

Using both Microsoft Excel and the Matlab toolkit Colour Engineering Toolbox [12] for cross validation, Z-scores with 95th

Figure 2. Colour images used in the re-evaluation of the psychometric experiment.

percent confidence interval were calculated from raw pair comparison data [13]. To follow the same procedure as the original paper [1], a binomial test was also done for every experiment.

Results and Discussion

In the psychometric experiments (see Figures 3 and 4), both the proposed algorithm (anisotropic Euclidean ΔE99) and our own greyscale reproduction performed significantly better than CIELAB L∗ when all images are considered.

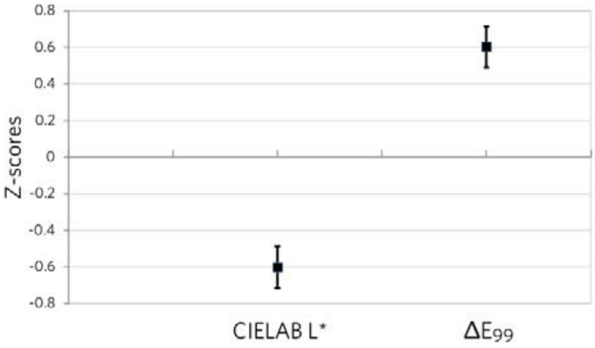

Figure 3. Z-scores with 95% confidence interval for the for the L∗ and the ΔE99 re-evaluated preference experiment.

For the comparison between the anisotropic Euclidean ΔE99) and manual method, the results (Figure 5) shows that the manual method is performing significantly better when all images are considered.

In line with the statistical evaluation done in the original paper, we also calculated the p-values from a binomial test for the three pair comparisons. In all cases we obtained as result $p < 104$, which shows that the evaluated methods performed significantly better than their pairs in the respective tests.

In the case of the anisotropic Euclidean ΔE99 reproductions, results show that when comparing images using the flip method, the advantages of the algorithm are visible to the observers and this thus there is a higher preference towards it (Figure 4). This makes sense, since the local details that are at times not perceived when comparing images side to side, are more noticeable in our experimental setup where the observers can freely flip between the images.

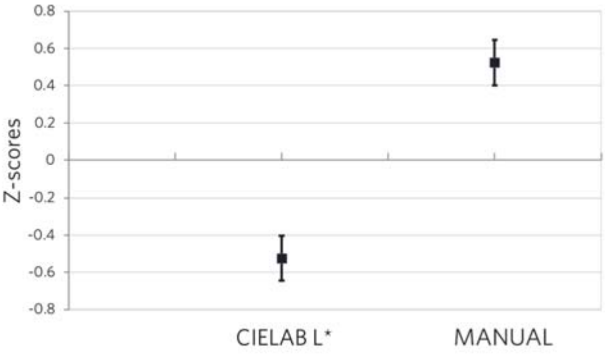

Figure 4. Z-scores with 95% confidence interval for the L⍰ and manual reproduction preference experiment.

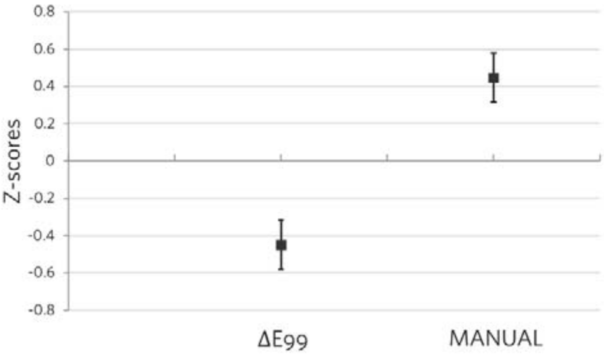

Figure 5. Z-scores with 95% confidence interval for the ΔE_{99} and manual reproduction preference experiment.

As for the manual reproduction, which did not use any spatial optimization to retain detail or local contrast, experiments showed that it performed significantly better than both algorithmic reproductions. These results show that observers not only rate more detailed images better, but also images that are pleasing. The main enhancement done to the images was to increase their contrast while retaining a pleasing photographic tonal reproduction, which also gives the perceived appearance of a sharper image. Retouching images in such way is a skill that colourists develop with practice, and that allows them to "improve" the images captured by cinematographers (a craft known as colour grading).

Conclusion

When improving the quality of greyscale reproductions, observers (more so those who are not specialized in working with images) can overlook the enhancements depending on the way in which the results are shown to them. The psychometric evaluation method used to evaluate the results of spatial algorithms is an important factor when testing the performance of such an algorithm. In this paper we evaluated colour-to-greyscale conversion algorithms in a subjective experiment where the observers could "flip" between the images.

While some algorithms go to great lengths to improve image quality, simple methods such as luminance maps can also produce better results. Such is also the case for the conversions done by a visual artist, whom can instinctively improve the perceptual quality of an image by judging the results "by eye".

References

[1] Ivar Farup, Marius Pedersen, and Ali Alsam. Colour-to- greyscale image conversion by linear anisotropic diffusion of perceptual colour metrics. In 2018 Colour and Visual Computing Symposium (CVCS), pages 1–6. IEEE, 2018.

[2] Ivar Farup. Hyperbolic geometry for colour metrics. Optics Express, 22(10):12369–12378, 2014.

[3] Dibakar Raj Pant and Ivar Farup. Riemannian formulation and comparison of color difference formulas. Color Re- search & Application, 37(6):429–440, 2012.

[4] Kaleigh Smith, Pierre-Edouard Landes, Joe¨lle Thollot, and Karol Myszkowski. Apparent greyscale: A simple and fast conversion to perceptually accurate images and video. In Computer Graphics Forum, volume 27, pages 193–200. Wiley Online Library, 2008.

[5] Hao Du, Shengfeng He, Bin Sheng, Lizhuang Ma, and Rynson WH Lau. Saliency-guided color-to-gray conversion using region-based optimization. IEEE Transactions on Image Processing, 24(1):434–443, 2015.

[6] G Cui, MR Luo, B Rigg, G Roesler, and K Witt. Uni- form colour spaces based on the din99 colour-difference formula. Color Research & Application, 27(4):282–290, 2002.

[7] Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: full-reference image quality assess- ment and the role of strategy. Journal of Electronic Imag- ing, 19(1):011006, 2010.

[8] Khai Van Ngo, Jehans Jr. Storvik, Christopher Andre Dokkeberg, Ivar Farup, and Marius Pedersen. Quickeval: a web application for psychometric scaling experiments. In Image Quality and System Performance XII, volume 9396, page 93960O. International Society for Optics and Photonics, 2015.

[9] Steven Le Moan and Marius Pedersen. Evidence of change blindness in subjective image fidelity assessment. In 2017 IEEE International Conference on Image Processing (ICIP), pages 3155–3159. IEEE, 2017.

[10] Steven Le Moan and Marius Pedersen. Measuring the effect of high-level visual masking in subjective image quality assessment with priming. In 2018 IEEE International Conference on Image Processing (ICIP), pages 3553–3557. IEEE, 2018.

[11] Gary G Field. Test image design guidelines for color quality evaluation. In Color and Imaging Conference, volume 1999, pages 194–196. Society for Imaging Science and Technology, 1999.

[12] Phil Green and Lindsay MacDonald. Colour engineering: achieving device independent colour, volume 30. John Wi- ley & Sons, 2011.

[13] Peter G Engeldrum. Psychometric scaling: a toolkit for imaging systems development. Imcotek press, 2000.

Author Biography

Aldo Barba received his MS in Colour Science and Industry – COSI from the Erasmus+ COSI Consortium which included the Norwegian University of Science and Technology, NTNU (2019). His research work was focused on gamut mapping for video and image quality evaluation methods.

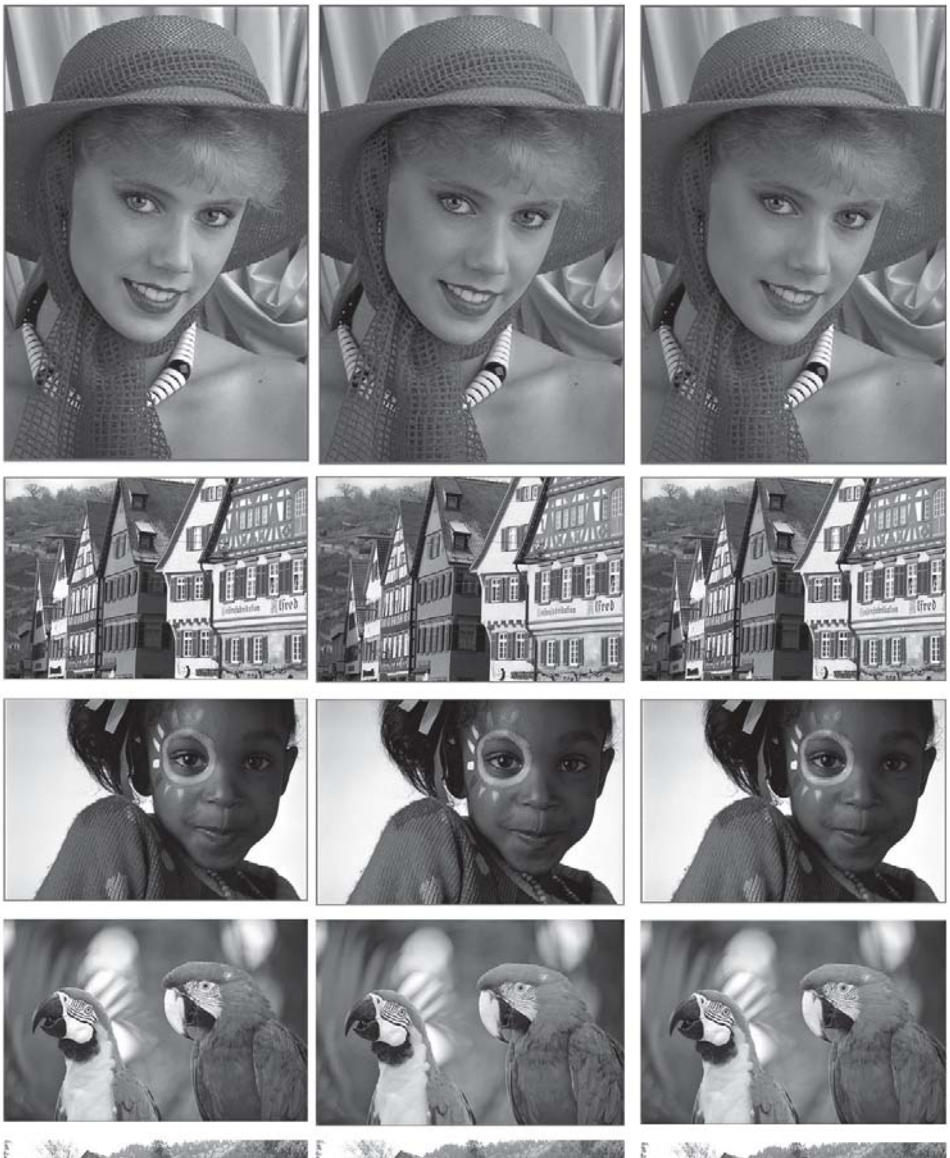

Figure 6. *Greyscale reproductions for CSIQ data set, from left to right: manual reproduction, CIELAB $L*$, $\Delta E99$.*

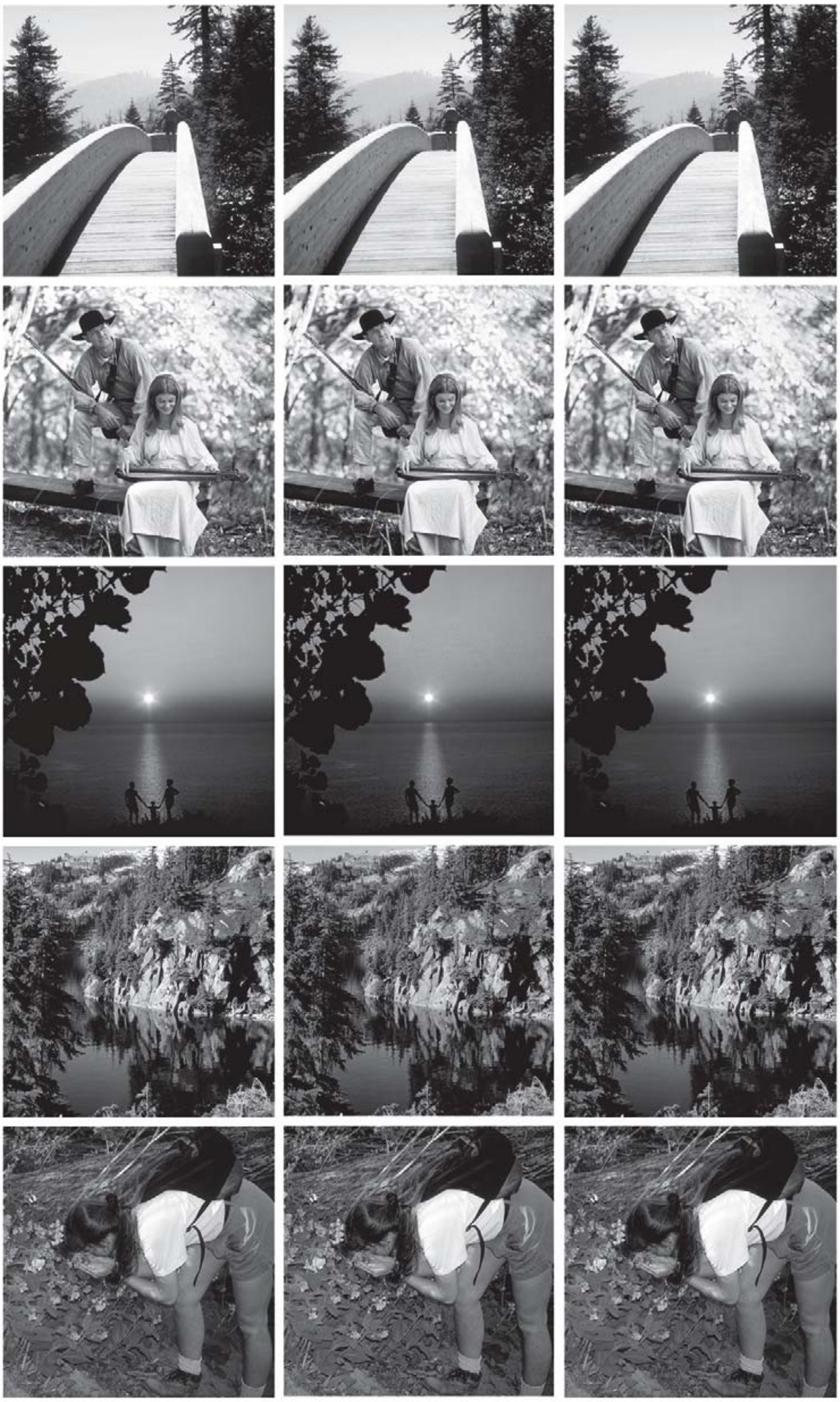

Figure 7. *Greyscale reproductions for Kodak dataset, from left to right: manual reproduction, CIELAB $L*$, $\Delta E99$*