"There are three principal means of acquiring knowledge . . . observation of nature, reflection, and experimentation. Observation collects facts; reflection combines them; experimentation verifies the result of that combination." ~ Denis Diderot (French man of letters and philosopher, 1713-1784)

Chapter 6

PSYCHOPHYSICAL EVALUATION OF THREE ALGORITHMS

In the previous chapter, a quantitative performance analysis of various algorithms was presented. This chapter contains a detailed description of the psychophysical experiments performed on outputs of several algorithms and an analysis of the results. Experiments were performed on still images as well as on video test sequences. Many of the algorithm implementations discussed in the pervious chapter were not integrated algorithms, focusing either on lightness adjustment, or on color enhancement, or simply contrast enhancement, but not all at the same time. Thus, it was not appropriate to include these algorithms in a single psychophysical experiment, as the end-results were very different. The experiments discussed in this chapter involves only three of the seven algorithms discussed, two Intel-proprietary algorithms CH and YO, and the proposed algorithm. All three algorithms attempt to enhance both color and contrast of the input images (or motion pictures).

This chapter starts with a discussion on display characterization process, which is essential for any psychophysical experiment involving display devices. Next, a paired comparison experiment conducted on several still images is described, followed by a similar experiment performed on video test sequences. Note that the experiments involved a Liquid Crystal Display (LCD) device, and so in the context of this chapter, a display mainly refers to an LCD.

6.1 Color Modeling of the LCD

For any display-based visual experiment, it is critical that the psychophysical images be transformed to colorimetric definitions accurately. This is only possible by accurately modeling, or characterizing, the inherent nonlinear nature of a computer-controlled display device. This nonlinearity is described by the optoelectronic transfer function (OETF), the relationship between the signals used to drive the display and the radiant outputs produced by the corresponding input. Determining this relationship is essentially a two-step process, namely, display calibration followed by a characterization.

6.1.1 Display Calibration

Device calibration is the process of maintaining the device with a fixed known characteristic of color response [**Sharma 2003**]. Many LCD display manufacturers build correction tables into the video card to convert the native luminance response curve (or gamma) into a desired response curve. It is advisable to turn off the built-in corrections if possible, and use the native gamma instead. As an example, the Display Calibrator application available in Mac OS goes through several steps in order to calibrate a display. These are: i) determining the display's native response, ii) selecting a target gamma (e.g. a linear gamma of 1, Mac standard of 1.8, PC standard of 2.2 or the native gamma), and iii) selecting the target white point (for example, native white point or D50 or D65). In the end, the calibration defines a display color profile, which includes the display gamut, the white point as well as the gamma. Figure 6.1 shows the report generated by Display Calibrator in Mac OS.

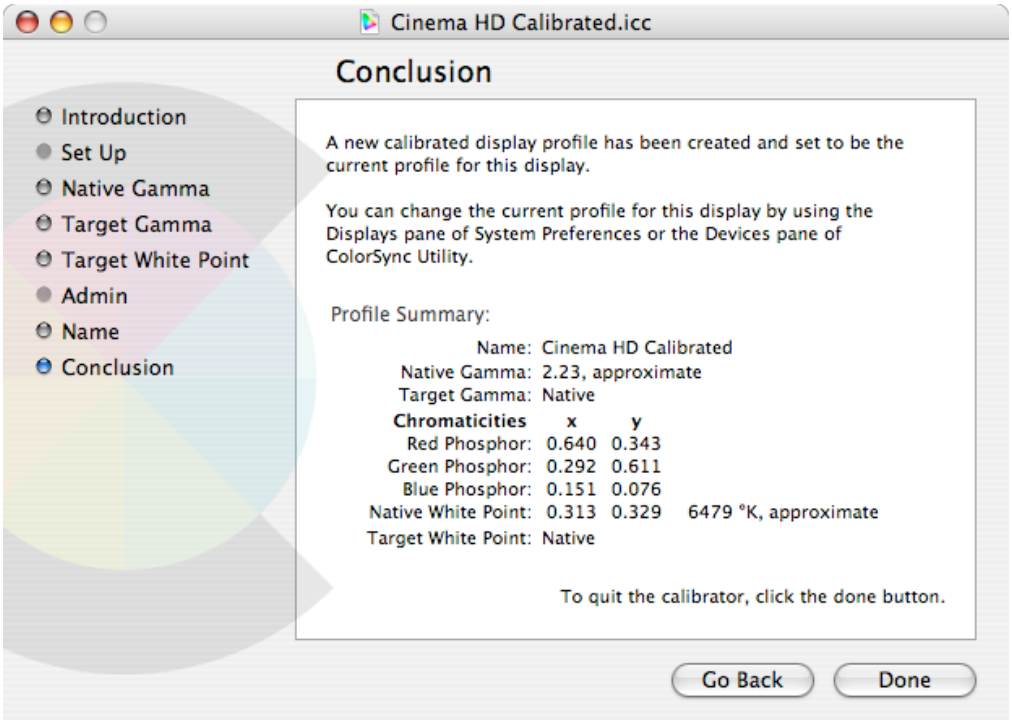

Fig. 6.1. Results of calibration using Display Calibrator in Mac OS

6.1.2 Display Characterization

Next, characterization is performed on the calibrated display. Characterization is a process that derives the relationship between device-dependent and device-independent color representations for a calibrated device. It is assumed that a calibrated device maintains the validity of the function, but the calibration process may need to be repeated from time to time to compensate for the temporal changes in the device's response and maintain it in a fixed known state [**Sharma 2003**]. Mathematical steps involved in the characterization process are described below. Discussion in this section is based on Day et al's work [**Day 2004**].

An effective way to represent the nonlinear characteristic in a computer-controlled display is to build one-dimensional look-up tables (LUTs) shown in Eq 6.1. These LUTs convert the original 0-255 RGB digital counts to linearized RGB values, thus defining the optoelectronic transfer functions for the three channels.

$$\begin{aligned} R &= LUT(d_r) \\ G &= LUT(d_g) \\ B &= LUT(d_b) \\ 0 &\le R,G,B \le 1 \end{aligned} \tag{6.1}$$

where d represents the digital counts and R, G and B are the radiometric scalars for the three channels, with values ranging between zero and unity.

The relationship between radiometric scalars and CIE tristimulus values is expressed by

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} X_{r,\max} - X_{k,\min} & X_{g,\max} - X_{k,\min} & X_{b,\max} - X_{k,\min} & X_{k,\min} \\ Y_{r,\max} - Y_{k,\min} & Y_{g,\max} - Y_{k,\min} & Y_{b,\max} - Y_{k,\min} & Y_{k,\min} \\ Z_{r,\max} - Z_{k,\min} & Z_{g,\max} - Z_{k,\min} & Z_{b,\max} - Z_{k,\min} & Z_{k,\min} \end{bmatrix} \begin{bmatrix} R \\ G \\ B \\ 1 \end{bmatrix} \tag{6.2}$$

where $X_{r,max}$, $Y_{r,max}$ and $Z_{r,max}$ are the maximum tristimulus values obtained from the r channel. Tristimulus values corresponding to other channels are expressed similarly. $X_{k,min}$, $Y_{k,min}$ and $Z_{k,min}$ are the black-level flare. The black-level flare terms are separated into a single column to form the 3x4 transformation matrix. The above primary transformation matrix can be optimized by minimizing CIE ΔE_{00} color difference for a dataset sampling the display's colorimetric gamut.

6.1.3 Experimental Setup

A 22" flat-panel Apple Cinema LCD was characterized and subsequently used in all psychophysical experiments. The display was controlled by a 4x2.5 GHz PowerPC G5 computer running Mac OS X 10.4, with a 2 GB DDR 2 SDRAM memory. The display had a maximum resolution of 2560x1600 pixels. As explained previously, the display white point and gamma were set to native values.

A Graphical User Interface (GUI) was designed to display a round patch in the middle of the screen. Red, green and blue color samples were generated each as 11 step ramp at equal interval. Uniform grey background was used throughout the experiment. White color was displayed and measured in order to determine the display white point. Display white point and CIE 10° observer data were used in all calculations. The tristimulus values of the colors were measured with an illuminance-type LMT colorimeter, which was interfaced to the computer. Measurements were taken in a completely darkened room. The tristimulus values were saved in a file.

Three one-dimensional Look-Up Tables (LUT) corresponding to 256 digital counts, describing optoelectronic transfer functions of each of the three channels were created from the colorimetric data, as shown in Figure 6.2. An optimized 3x4 transformation matrix (Eq 6.2) to convert the digital count to tristimulus values was also created. Nonlinear optimization was used to minimize ΔE_{00} between the estimated and measured color patches. All the matrix coefficients were estimated simultaneously during the optimization process, changing the LUTs dynamically during each iteration.

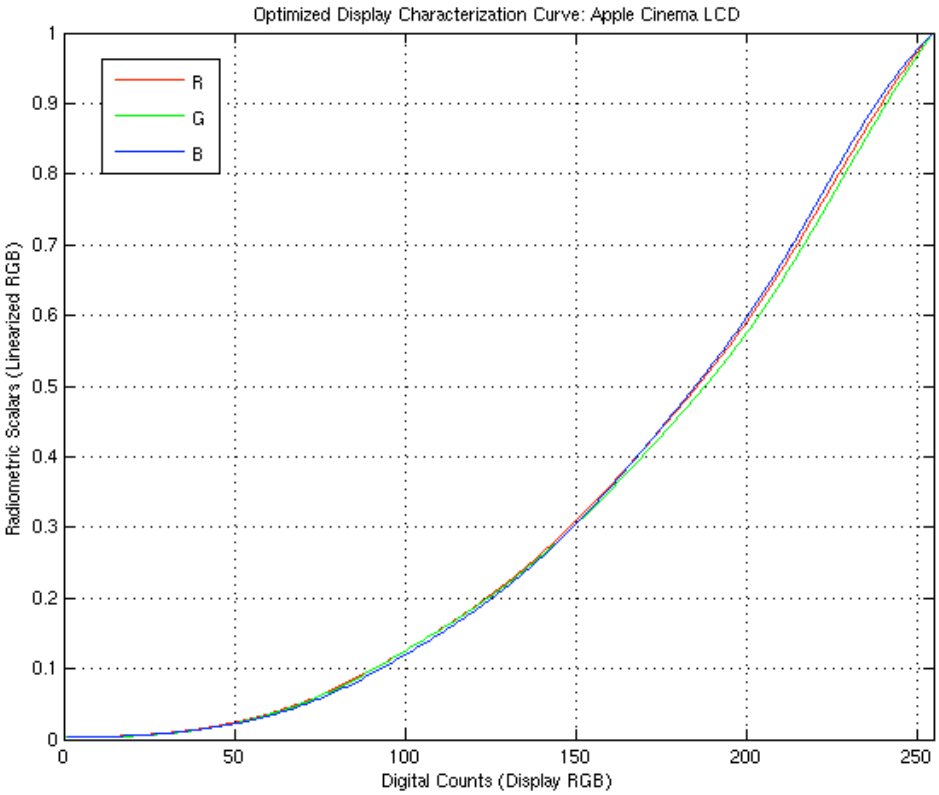

Fig. 6.2 Optimized Display Characterization Curve: Apple Cinema LCD

6.2 Psychophysical Experiments

Image quality is an integrated perception of the overall degree of excellence of an image [**Engeldrum 2004**]. Psychometric scaling is widely used in the imaging field for obtaining scale values of image quality and the "nesses", or its attributes. The process of scaling is not always straightforward. Several factors need to be taken into account during sample selection, choosing observers, formulating task instructions and finally, presenting and viewing the samples. The following subsections describe various aspects of the experimental design as well as an analysis of the results.

6.2.1 Experimental Goal

A key aspect of this research was to conduct subjective tests on the outputs of different algorithms. This was considered a potentially important contribution to the field since most of the development efforts published so far lack a systematic subjective assessment of the image enhancement algorithms. The main objective of conducting these psychophysical image quality experiments was to compare the performance of the new algorithm against some benchmark. In this case, the benchmarks were two image enhancement algorithms developed within Intel.

6.2.2 Software for Psychophysical Experiments

SiQ (pronounced sai-que), a software tool previously developed by the author was used for designing and executing the experiments as well as for analyzing the results. The software was developed in Matlab environment and had a graphical user interface. While the presentation of the trials for the still image experiment was handled by SiQ itself, some additional processing was required for the video experiment as described later.

6.2.3 Algorithms Evaluated

Apart from the new image enhancement algorithm, two Intel-proprietary algorithms were included in this test. Throughout this report, these algorithms are referred to as **CH** and **YO**. The new algorithm has been referred to as **NA**. The original images were also included in this experiment. These are referred to as **OR** in the figures. So there are four versions of each test image.

6.2.4 Test Images

Fifteen still images were included in the first psychophysical experiment. Each image had a size of 920x720 pixels.

Selection of the test images is extremely important for an unbiased evaluation of image enhancement algorithms. The following few pages contain the details of the test images used in this experiment. Note that for various color management and color appearance issues, images reproduced in this document will not have the same appearance as the actual images viewed on a characterized display. The contrast of the images in this document will generally appear a lot higher because of the reduced size. For the same reason, the output images are not included here. Various characteristics of output images generated by different algorithms have been discussed in the previous chapter.

Test Image 1

Critical aspects: Closely knit shiny beads of different colors.

Goal: To enhance color and contrast in such a way that individual beads as well as the specular reflections on them are distinctly visible

Test Image 2

Critical aspects: People's faces

Goal: This image is a little out-of-focus, and needs contrast enhancement. Because of the image content, the color enhancement has to be relatively subtle.

Test Image 3

Critical aspects: The animals, the blue jacket

Goal: Haziness due to the dust in the air needs to be preserved while enhancing the contrast

Test Image 4

Critical aspects: The colors, creases and/or the design in the dresses, skin tone, the green forest in the background

Goal: The skin tone and the greenery in the trees must not undergo strong enhancement. The man and the woman are a little blurred due to motion

Test Image 5

Critical aspects: Green grass outside the tunnel, the joints and the graffiti on the tunnel wall, the overhead fixtures

Goal: Preserve the high dynamic range in the scene; avoid introducing noise in the dark areas of the image; avoid strong enhancement of the grass

Test Image 6

Critical aspects: Colors of different vegetables, water droplets on the tomato, image contrast, particularly on the cauliflower, the broccoli and the corn

Goal: Avoid turning achromatic colors into chromatic ones while enhancing color and contrast of the vegetables

Test Image 7

Critical aspects: Same as the previous image. This is a low-contrast version of test image 6.

Goal: Increase the overall contrast (in addition to the stated goals for test image 6)

Test Image 8

Critical aspects: The bricks on the buildings, the stairs, the texts, the objects in the shadowed areas, the sky

Goal: White letters as well as the objects in the shadowed area should be more visible, the brick pattern should texts in general should look sharper

Test Image 9

Critical aspects: Intricate artwork on the walls, creases on the dresses of the people, bright light spots above the torchieres , light and shadow on the floor

Goal: enhance color and contrast while preserving the intricacies of the sculptured walls

Test Image 10

Critical aspects: Skin tone, hair on woman's forehead, her earrings, the blue denim shirt

Goal: Enhance color and contrast while avoiding strong color enhancement of the skin tone

Test Image 11

Critical aspects: color of the flowers, green plants, stone walls, dark areas under the big trees

Goal: enhance color and contrast of the flowers and green plants while preserving the naturalness; enhance the contrast of the dark areas so that shadowed objects are more visible

Test Image 12

Critical aspects: Different skin tones, colors on the dresses, text on the green badge

Goal: Avoid strong enhancement of different skin tones while enhancing color and contrast

Test Image 13

Critical aspects: Face of the man on the left, the text

Goal: Avoid enhancing the noise in the original image while boosting the contrast; avoid strong enhancement of the skin tone

Test Image 14

Critical aspects: The colorful parts in the balloon

Goal: Avoid making the noise in the original image more apparent while enhancing the color and contrast

Test Image 15

Critical aspects: LEDs and the overall contrast in this night scene

Goal: Avoid making the noise in the original image more apparent while enhancing the color and contrast

6.2.5 Test Movie Sequences

A second psychophysical experiment was performed on four video test sequences. As in the first experiment, outputs from the proposed algorithm, CH, YO, as well as the original sequences were included. The number of sequences had to be limited to four because of various reasons. Firstly, sequences appropriate for color and contrast enhancement were not readily accessible at the time of this research. Secondly, outputs from the proprietary algorithms CH and YO had to be obtained through the sponsor since IP issues were involved. Finally, the duration of the experiment had a practical limitation on the number of test sequences. The sequences had durations between 7.5 seconds and 10 seconds.

The movie sequence Avia had a resolution of 854x480. All other movie sequences were cropped to the same resolution to maintain uniformity. Having a resolution significantly larger than this for any sequence caused a playback and synchronization problem, in spite of using one of the most powerful computing resources available in the laboratory (configuration discussed earlier under experimental setup). The file sizes of the frames in these sequences varied from 1.1 MB to 1.3 MB. Processing did not change the file size appreciably, as expected.

A description of the video sequences follows.

6.2.5.1 Movie Sequence “Avia”

Description: This was a TIFF image sequence of 224 uncompressed images. This was not a continuous sequence, but rather selected from different locations in a long sequence. The duration of the clip was approximately 7.5 sec.

Content: This was mainly a restaurant scene starting with a close-up of desserts, and then panning on to the man and woman talking, then a close-up shot of fresh salad. Next, the clip shows a man talking to another man and a woman sitting in front of him. Finally, there is a close-up shot of a woman sitting against a black background with a color chart. Several frames of the sequence is shown in Figure 6.3.

Fig. 6.3 Different clips from the sequence Avia

Critical aspects: Various colors present in the content are important from the perspective of this experiment. Examples are the strawberries in the dessert, the wine glasses and other materials on the table, the skin tone, the colors of the carrot and other vegetables, the red shirt of the woman sitting in the couch and the color charts in the last several frames.

In several frames, contouring resulting from the compression led to challenging content for image enhancement. For example, the background in the first and third example frames shown in Figure 6.3 or the wall in the 4th example frame showed contours in the input image. If an algorithm makes these contours more visible while enhancing the contrast, it may result in reduced perceived video quality.

6.2.5.2 Movie Sequence "Calendar"

Description: This was a bitmap image sequence of 300 uncompressed images. The original size of 1280x720 was appropriately cropped so that the frames contained enough color even after cropping. The duration of the clip was approximately 10 sec.

Content: The clip shows a black and white stripe pattern along with some content with colors and a part of a calendar showing dates, as shown in Figure 6.4. The whole content rotates slowly throughout the duration of the clip.

Critical aspects: There is not a lot of color in this sequence. The stripe pattern or the numerals in the calendar are important for contrast enhancement evaluation and any resulting artifacts. The image is inherently noisy, so contrast enhancement might boost the noise as well, leading to poor picture quality. Even though noise was not included as a parameter in the psychophysical experiment, it is difficult for the observers to fully discount the effect of noise enhancement.

Fig. 6.4 Clips from the sequence Calendar

6.2.5.3 Movie Sequence “Vintage Car”

Description: This was a bitmap image sequence of 300 uncompressed images. The original size of 1280x720 was appropriately cropped so that the main object of interest in the clip, the vintage car in motion, was included in all frames. The duration of the clip was approximately 10 sec.

Content: The sequence shows a blue-colored vintage car approaching from a distance through a wooded area full of green foliage. The last several frames contain a close-up shot of the car and the man driving it. Figure 6.5 shows some representative frames.

Critical aspects: The foliage and the car are the main objects of interest with regard to color. In the close-up shots, the skin tone of the man and the color of the car are of importance. As in the Calendar sequence, there is noticeable noise in the input movie sequence, which is likely to get enhanced during contrast enhancement, unless a noise reduction module is included in the algorithm.

Fig. 6.5 Clips from the sequence Vintage Car

6.2.5.4 Movie Sequence "Walking Couple"

Description: This was a bitmap image sequence of 250 uncompressed images. The original size of 1280x720 was appropriately cropped so that the man and the woman walking down the wooded path were included in all frames. The duration of the clip was approximately 8.5 sec.

Content: The sequence shows a man and a woman walking together amidst the woods in the backdrop of trees. The man is wearing a bright yellow shirt and a colorful tie, while the woman is wearing a pink shirt with patterns on it. Two frames from the sequence are shown in Figure 6.6.

Critical aspects: The colors on the clothes are the most prominent features in this sequence, apart from the thick foliage in the background. The skin tone is also important.

Fig. 6.6 Clips from the sequence Walking Couple

6.2.6 Viewing Conditions

All images and video sequences were run through the Lookup Tables obtained from display characterization before displaying on the LCD screen. The experiments were performed in a completely dark room. The observers maintained a distance of around 30 inches from the screen.

6.2.7 Observers

A total of 25 color normal observers participated in each psychophysical experiment involving still images and the video test sequences. While both naïve and experienced observers were included in the experiments, no observer was familiar with the algorithms or the technology variables. It is important to not include such observers as they might have significantly different image preferences than the average observers [**Engeldrum 2001**]. Most of the observers were students and staff at the Munsell Color Science Laboratory, Rochester Institute of Technology. It is worthwhile to note that repeated empirical observation showed that experts and non-experts

judge image quality similarly when the task is application-independent, resulting in an image quality scale that is more "absolute" [**Engeldrum 2004**].

Before the actual experiments started, color vision of the observers was tested with some of the test plates from Ishihara Pseudoisochromatic test 24-plate [http://www.toledo-bend.com/colorblind/Ishihara.html]. This screening test was done by SiQ itself.

6.2.8 Experimental Method for Still Images

Since the main objective of the experiment was to generate an interval scale of image preference for the algorithm outputs, the method of paired comparison was determined to be the best method to use. This is one of the most common experimental techniques to quantity image quality [**Wu 1998**]. Note that this method generates a one-dimensional scale. In other words, we assume that the variability in the observers' responses can be fully expressed in a single dimension.

Every pair of the images was presented to the subject in a unique random order chosen by the software. The relative position of the images on the display screen was also randomized. The same pair of samples was presented only once. There were 90 observations in all (a pair can be chosen from 4 versions of a given image in ${}^4C_2$ or 6 ways, and there were 15 test images, so 6x15 = 90). Each session was completed in approximately 25 minutes on an average.

The following instruction was given to each observer:

Thank you for participating in our study.

There are 15 images. For each image, there are 4 versions. This experiment has 90 observations in all. In each observation, two versions of the same image will be displayed on the screen. Choose the image that you prefer. If you prefer the left image, click on the box displayed below the left image. If you prefer the right image, click on the box displayed below the right image.

There is no time limit, and there is no right or wrong answer. We are seeking your opinion. If you have any questions, please ask the test administrator at this point.

6.2.9 Experimental Method for Video Test Sequences

Similar to the still image experiment, the method of paired comparison was also used in the experiment involving video test sequences. Two movies of the same content were shown simultaneously, one at the top of the window and one at the bottom, playing in a continuous loop. The observers were allowed to take as long as needed before deciding on their preference. Additional processing was needed to prepare the video for this test as described below.

Step 1: Exporting Original and Processed Image Sequences As Quicktime Movies

The original and processed image sequences were first converted into movies using QuickTime Pro for Mac. Image compression was applied while generating the movies. While this was not a preferred method, uncompressed movies could not be played in the computer without experiencing playback problems. Note that two movies had to be played simultaneously, so synchronization was also an issue with uncompressed movies. The image compression reduced

the file sizes, and thus reducing the memory requirement during playback. H.264 image compression method with medium quality setting and a frame rate of 29.97 frames/sec was used. Every 24th frame was designated as a key frame in the compression process.

Step 2: Combining Clips in Pair Using SMIL Scripts

Once the movie clips were ready, SMIL (Synchronized Multimedia Integration Language) scripts were written to generate combined movie clips with every possible combination [**SMIL 2005**]. There were four versions for each movie clip (original, new algorithm and two Intel proprietary algorithms), which could be combined in ${}^4C_2$ * 2 or 12 ways, since a given version could be played either on top, or on bottom. Thus, for four input movie clips, there were 48 combined movies. With the chosen resolution, the combined movie clip could be played without any synchronization problem.

Step 3: Loading Combined Movie Clips in A Web Browser

In the final step, HTML script was written to load a combined movie clip onto a web browser. In this case, the browser was Safari (for Mac). The browser parameters, including the background color, exact position of the movie in the browser window were set within the script. The whole background was set to gray. This script was saved as a webpage. During the experiment, SiQ edited the script to insert the name of the appropriate movie according to the predefined trial sequence.

The following instruction was given to each observer:

Thank you for participating in our study.

In each trial, two versions of a given clip will be randomly picked by the software and will be presented to you in a browser. The movies will continuously play in loop. Determine which clip has the highest OVERALL picture quality in terms of color, contrast and sharpness. IGNORE NOISE IN THE MOVIES.

Once you have decided which clip to choose, click on the appropriate button (top or bottom) in the input window based on your choice. If you accidentally close the browser, you can reload the current trial by pressing the reload button. Please DO NOT resize the browser at any time.

There is no time limit, and there is no right or wrong answer. We are seeking your opinion.

To get started, click OK and then click anywhere in the browser. If you have any questions, please ask the test administrator at this point.

6.3 Results and Discussion

Data from complete pair wise comparisons were analyzed using Thurstone's Law of Comparative Judgment Case V [**Thurstone 1927**] in order to create an interval scale of overall image preference.

6.3.1 Thurstone's Law of Comparative Judgment

Following are the hypotheses behind Thurstone's Law of Comparative Judgment (Thurstone 1927):

1. Each stimulus gives rise to a discriminal process, which has some value on the psychological continuum of interest.
2. Due to momentary fluctuations (internal fluctuations occurring within or between observers), the value of a stimulus may be higher or lower on repeated presentations. The distribution of this fluctuation can be characterized by a normal distribution.
3. The mean and standard deviation of the distribution associated with a stimulus are its internal scale values and discriminal dispersion, respectively.
4. The distribution of the difference between two stimuli is also normally distributed and is a function of the proportion that one stimulus is chosen as greater than the other.
5. The difference in scale values, R, between two stimuli, i and j, is:

$$R_i - R_j = z_{ij}\sqrt{\sigma_i^2 + \sigma_j^2 - 2r_{ij}\sigma_i\sigma_j} \tag{6.3}$$

where R_i and R_j represent the scale values of stimuli i and j, σ_i and σ_j, are the standard deviations of the respective discriminal dispersions, r_{ij} is the correlation between the two discriminal

processes, and z_{ij} is the normal deviate (the z-score) corresponding to the proportion of times stimulus *j* is judged is judged greater along the psychological continuum than stimulus *i*.

Eq (6.3) can be simplified based on certain assumptions:

1. The evaluation of one stimulus along the continuum does not influence the evaluation of the other in the paired comparison ($r_{ij} = 0$).
2. The dispersions are equal for all stimuli ($\sigma_i = \sigma_j$).

Accordingly, following equation is obtained:

$$R_i - R_j = z_{ij}\sqrt{2} \tag{6.4}$$

6.3.2 Confidence Interval

In terms of interval scale unit, standard deviation σ can be expressed as

$$\sigma = \frac{1}{\sqrt{2}} \tag{6.5}$$

Standard error of the scale value is

$$\frac{\sigma}{\sqrt{N}} = \frac{1}{\sqrt{2N}} = \frac{0.707}{\sqrt{N}} \tag{6.6}$$

Where N is the number of observations per pair, which is equal to the number of observers if each observer views a trial a single time.

The confidence interval for the paired comparison data can be expressed as

$$\mu = \pm F \frac{\sigma}{\sqrt{N}} \tag{6.7}$$

Here F is a function of the Level of Confidence (LOC). F is derived from the Normal Curve.

With 95% LOC, F = 1.96. So, the confidence interval is

$$\mu = \pm 1.96 \frac{0.707}{\sqrt{N}} = \pm \frac{1.38}{\sqrt{N}} \tag{6.8}$$

Note that the confidence interval does not depend on the number of observations – this is the weakness of this statistical metric.

Since there were 25 observers in each experiment, so the confidence interval was 0.276. Note that the confidence interval computed using above equations is the same for all images/sequences in a given experiment.

6.3.3 Interval Scale Plots: Still Image Experiment

An analysis of the experimental data leads to a separate interval scale for each test image (i.e. 15 interval scales in all). An average of these 15 interval scales is presented in Figure 6.7. Strictly speaking, we should not combine different interval scales since each image is a different psychophysical experiment and there is no common anchor point in these scales so that they can be tied together. Another issue with using this figure is that it does not reflect the fact that the observer preferences are strongly dependent on the image content, as will be clear shortly. However, this figure gives an impression on how each algorithm performed for all the images on an average. Existing algorithms CH and YO performed equally well, while the new algorithm

probably performed slightly better. All three algorithms generally performed better than the original. The error bars in this figure signify confidence intervals, or statistical uncertainty. Note that the zero does not have any specific meaning in the interval scale plot, only the relative positions are important. In other words, we can add any arbitrary constant to the interval scale values without affecting the results.

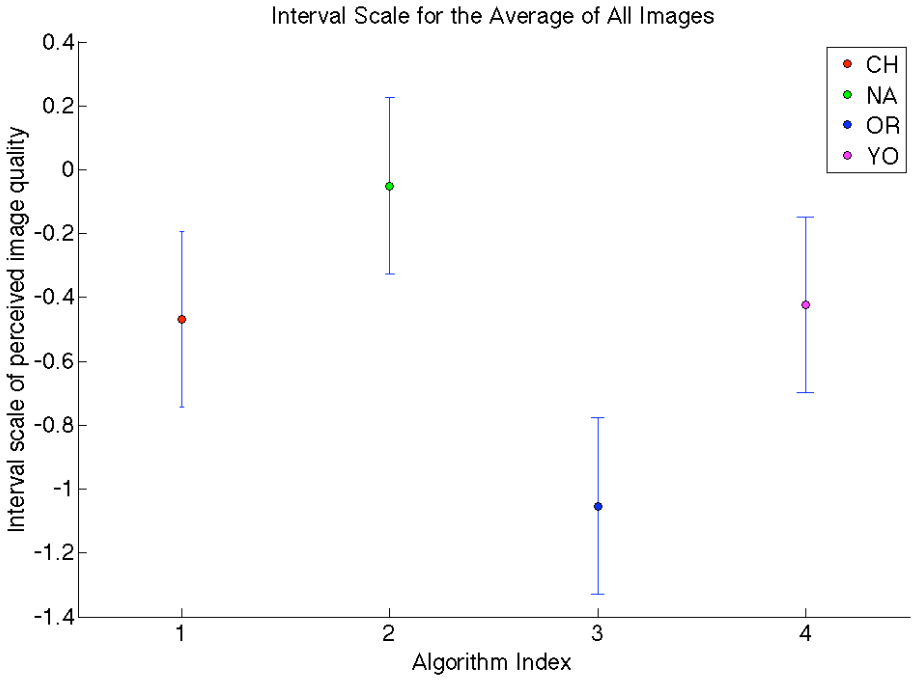

Fig. 6.7. Interval scale for the average of all images

The implication of the confidence interval is, based on this figure alone, we cannot make an inference that the new algorithm will perform better than the existing ones under all circumstances. This is more evident when we look at the result for all test images, as shown in Figure 6.8.

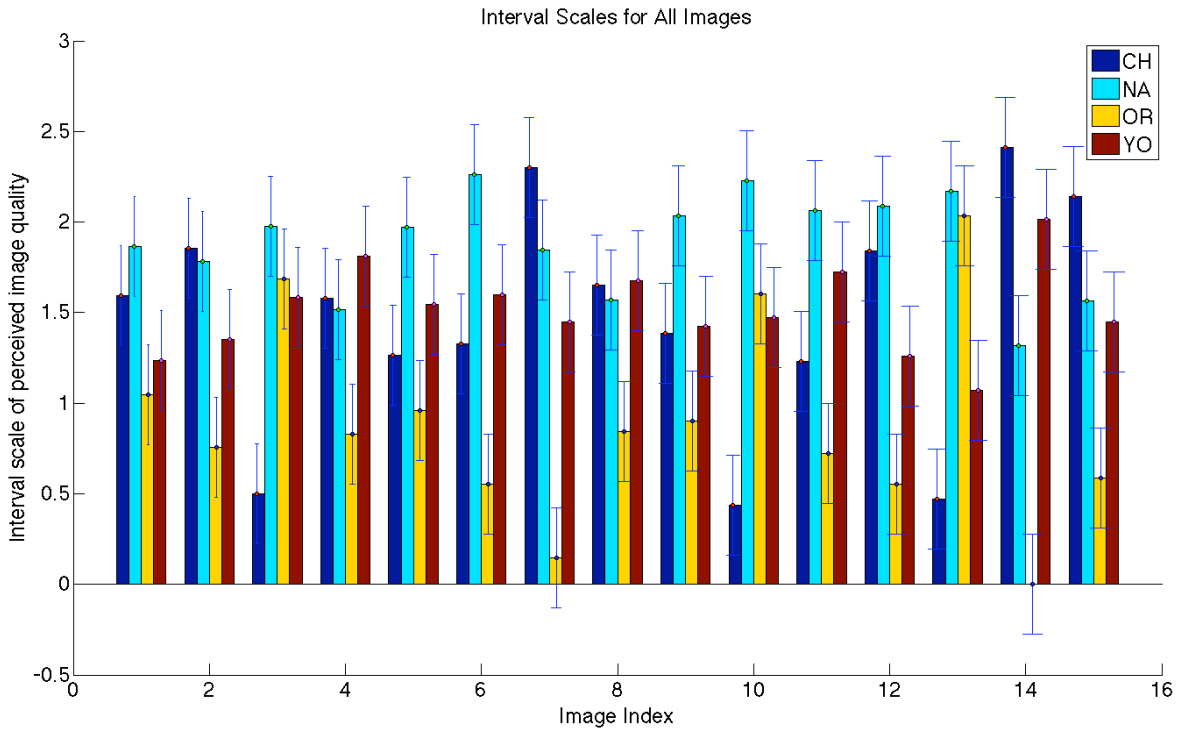

Fig. 6.8. A summary of interval scales for all test images

In the above figure, the interval scales have been shown for all 15 test images. Each bar corresponds to one version of the image (three algorithm outputs, or the original). Evidently, not a single algorithm was preferred for all these images. For many images, difference in the interval scale values for two or more algorithms is statistically not significant. Less is the overlap between two error bars, the more statistically significant is the corresponding interval scale difference. Table 6.1 summarizes observer preferences that can be considered statistically significant.

Table 6.1. Ranking Table for the performance of different algorithms in the still image experiment

Algorithm	Ranked #1		Ranked #3 or #4	
	Images	No of times	Images	No of times
CH	7, 14, 15	3	3, 10, 11, 13	4
NA	5, 6, 9, 10	4	14	1
OR	-	0	1, 2, 4, 5, 6, 7, 8, 9, 11, 12, 14, 15	12
YO	-	0	2, 12, 13	3

Note that even though algorithm YO output was not ranked #1 for any image while CH output was ranked #1 thrice, the overall ratings for the two algorithms are comparable (from Figure 6.6), as the performance of CH was worse for several images (e.g. image 10 and 13).

Comparatively, the new algorithm has performed consistently well. Its outputs were ranked #1 for four test images, and it performed significantly worse compared to the other algorithms only once (test images 14). In comparison, CH and YO were less consistent. They were ranked among the worst two versions 4 and 3 times respectively.

As expected, for most of the test images the algorithm outputs were preferred over the originals.

Note that, this result is specific to the set of test images used in this experiment. For a different set, this result might vary. However, the set used here includes a wide variety of image content and thus, the experimental results give a fair idea about how these algorithms might perform under different circumstances.

6.3.4 Interval Scale Plots: Video Experiment

As in the still image experimental analysis, the software SiQ generated interval scale for the average of all clips (Figure 6.9) as well as for individual movie clips (Figure 6.10). There is more ambiguity in these results than in the results obtained from the still image experiment. The difference in the interval scale of perceived picture quality for the three algorithms is statistically not significant.

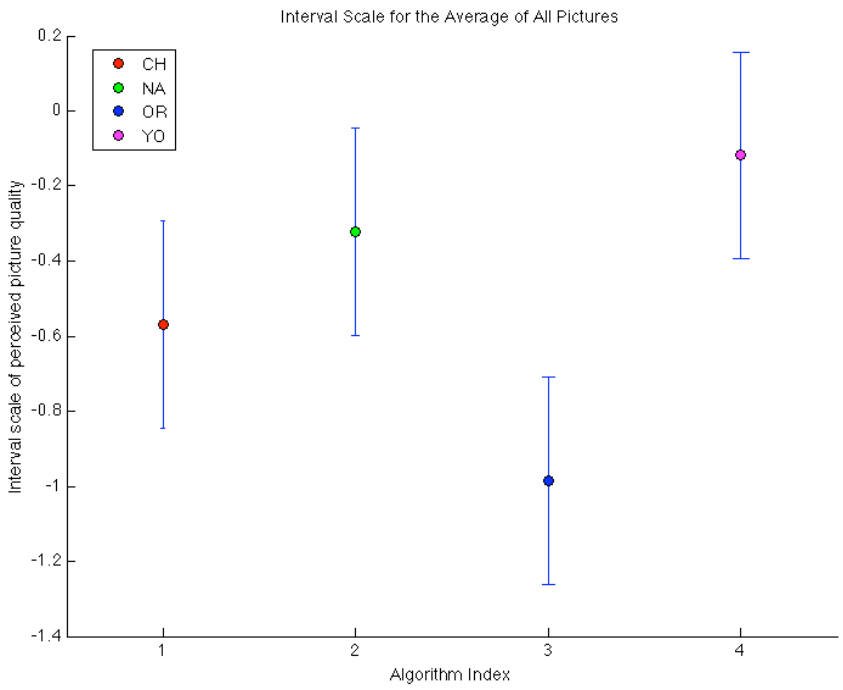

Fig. 6.9 Interval scales for the average of all clips

However, we can draw some general conclusions from the individual interval scales shown in Figure 6.10. Outputs of all three algorithms did better or similar to the original. This is more obvious for algorithms NA (proposed) and YO, whose performances were very similar for all four clips. Performance of CH was noticeably better than the other algorithms for the Calendar clip, but for other clips, it did not quite improve the perceived picture quality compared to the original.

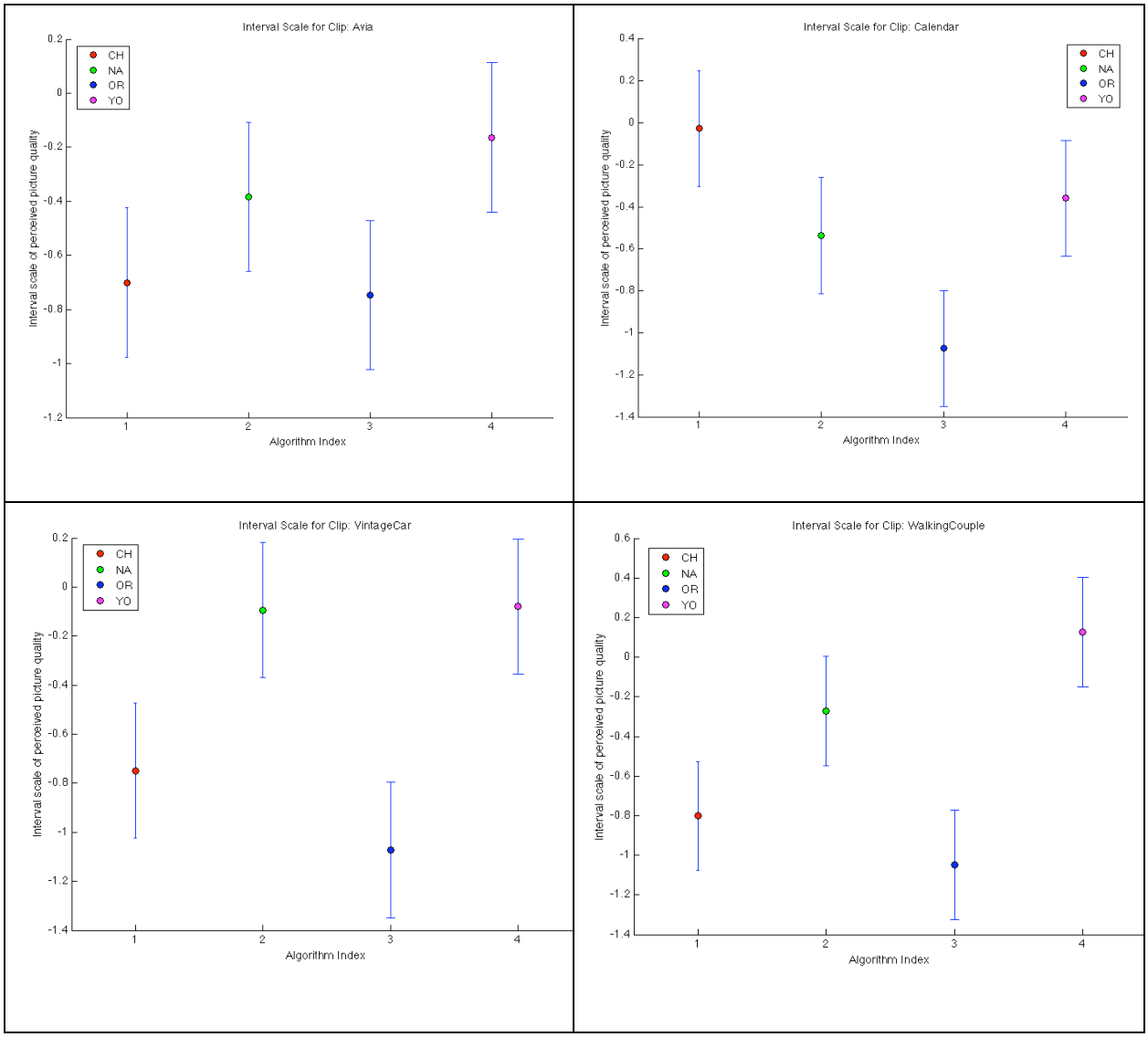

Fig 6.10 Interval scales for the four movie clips

Clip Avia shows more statistical uncertainty than the other clips. This is likely due to the image content. The change in the subject matter in this clip was rather fast, leaving observers with less time to discern perceptual difference between various outputs.

The fact that NA and YO showed a consistent performance across the clips is also evident in Figure 6.11, which shows the interval scales for all four clips and for all four versions of them.

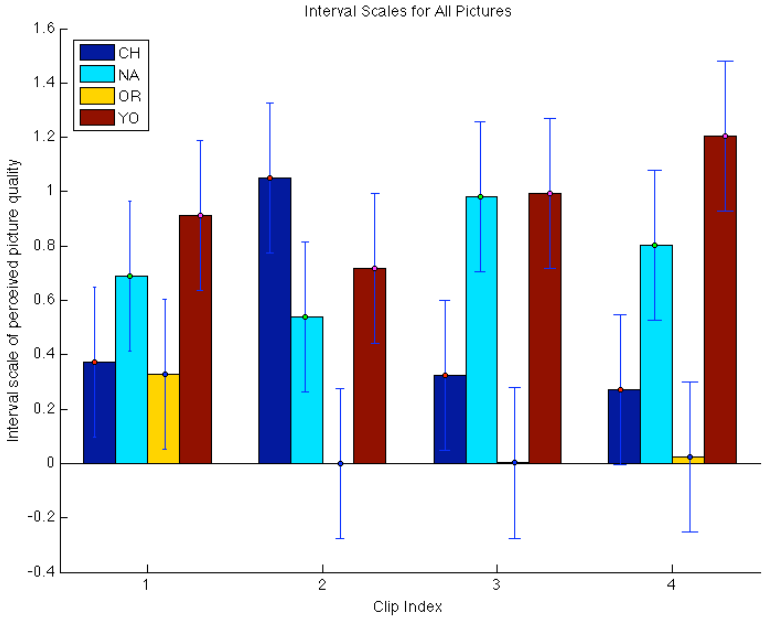

Fig. 6.11 Summary of interval scales for all movie clips

The somewhat inconclusive results from the video experiment can be attributed, at least partially, to the lack of appropriate image content. Movie clips available for this experiment were not probably very appropriate for evaluation of color and contrast enhancement aspects of different algorithms. The effect of color/contrast enhancement did not turn out to be very apparent in many cases, leading to the confusion or ambiguity in observer data. As explained earlier, higher

resolution clips could not be used in this experiment because of playback issues. Image compression had to be applied to reduce the memory requirement during playback. These imposed major limitation on the experimental setup. Reduced resolution has a direct impact on perceived contrast. The process of image compression arguably added some unknown variables into the processing chain, which might have affected the perceived picture quality.

6.3.5 Inference from the Results

The results from the psychophysical experiments, particularly the one involving still images, indicate that the new algorithm is performing well in most of the cases, although there are areas where it needs improvement. When an image is inherently noisy, the contrast enhancement in the algorithm is causing that noise to be enhanced as well, leading to a poor quality image. This is evident in cases of test image 14 and 15. This problem can be circumvented by introducing a noise detection mechanism that will prevent any contrast enhancement in case of noisy image contents. Alternately, noise detection or even noise suppression mechanism can be a part of the video processing chain, and the noise information can be provided as an algorithm input.

Even though skin tone enhancement worked reasonably well in the proposed algorithm, as evident in case of test image 10, a separate skin tone detection mechanism may need to be incorporated to improve the performance of the algorithm for this type of image contents.

With respect to processing image sequences, the results were not markedly different from the still images. No major temporal artifacts were noticed in any of the three algorithms included in the visual experiments. However, since the new algorithm relies on cumulative distribution

function to determine the intensity enhancement required in a given frame, chances are in some specific cases, there will be a noticeable shift in color from one frame to the next. Some sort of temporal processing should be included as a safeguard against that kind of situation. However, this problem was not observed in the test clips included in the visual experiment.

Overall, the performance of the new algorithm is quite promising. Further development can make this algorithm more successful as an automatic image enhancement method.