Evaluating Digital Film Look

Jurgen Stauder, Laurent Blondé, Thomson R&D, Cesson-Sevigné, France; Alain Trémeau, LIGIV, University of St. Etienne, France; Joshua Pines, Technicolor, Burbank, California, USA

Abstract

In cinematographic post production, digital processing of images - called Digital Intermediates (DI) – replaces more and more the traditional film workflow. Digital post production requires the preview of DIs with a reproduction of colors, dynamics and resolution comparable to the final film projection. This paper addresses the colorimetric reproduction by a color management approach and specifically develops a series of subjective tests to evaluate reproduction quality. Three subjective tests are developed. A first test – called Double-Stimulus Continuous Relative Quality Scale method (DSCRQS) - derived from the ITU-R BT.500-10 DSCQS test method allows non-biased, temporal digital versus film comparison. A second test, derived from ITU's DSIS and SDSCE tests, allows more sensible side-by-side comparison, but is biased. Finally a third test is introduced as a free in-depth side-by-side comparison to collect expert's comments. The test are based on a number of principles such as use of real film content, consideration of use cases and limitation of bias. A first test run of a test of second type was successfully applied to measure the effect of a change of a film projector's bulb in a color correction theatre.

Introduction

In cinematographic post production, digital processing of images - called Digital Intermediates (DI) – replaces more and more the traditional film workflow. Digital post production requires the preview of DIs with a reproduction of colors, dynamics and resolution comparable to the final film projection. This paper addresses the colorimetric reproduction by a color management approach and specifically develops a series of subjective tests to evaluate reproduction quality.

The paper is structured as follows. First, the color management approach for DI process is presented. Then, film look creation and its evaluation are discussed. Finally, results of a first test run of one of the proposed tests is presented.

Color Management Approach

Figure 1 shows the color management approach in the DI process.[1,5] The target device is a digital projector at bottom left side. (It can be replaced by any other display device.) The projector is set to film look using a Look Up Table (LUT) stored in the color box and applied to the RGB signal. The LUT is based on a forward device model for film and an inverse device model for the projector. The forward device model describes mathematically the film printing and projection chain. The inverse device model describes the digital projection process.

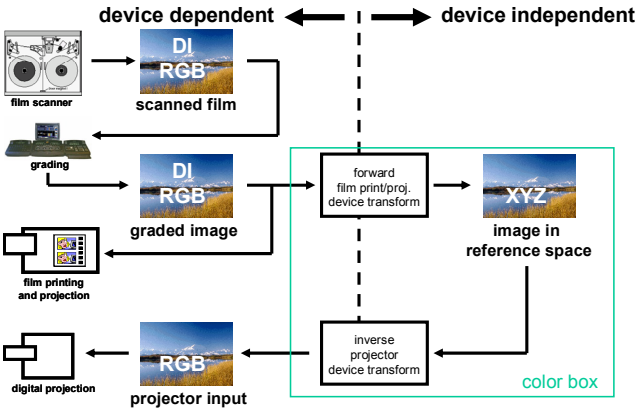

Figure 1. Calibrated DI workflow

The device models are established by device characterization. Device characterization is the process of building a device model from measurements. For a display device, measurements are the set of RGB input values and measured XYZ values of displayed colors. RGB are device dependent color input signals of the considered device, XYZ are values from CIE 1931 XYZ space for 2-degree observer. Device characterization generates a device model offering the following functionalities:

1. Interpolation of XYZ output values from given RGB input values (forward model)

   $$(X,Y,Z)^T = f_{XYZ}(R,G,B)$$

   where $(\ )^T$ is transpose operation.

2. Interpolation of RGB input values from given XYZ output values (backward model)

   $$(R,G,B)^T = f_{RGB}(X,Y,Z).$$

Device characterization is not the focus of this paper. For the experiments, a proprietary solution was used.

Characterization errors can be one source for errors in digital "film look". They may be caused by measurement errors (noise), dynamic flare (mutual refelection between screen and theater surfaces), imaging chain errors (spatial distortions, temporal variations) and inherent characterization errors (model errors, gamut mapping).

Film Look Creation

The target device is set to film look using Look Up Tables (LUTs). RGB-RGB LUTs are built from a forward film chain device model (RGB to XYZ) and an inverse digital chain device model (XYZ to RGB) such as shown in Figure 1.

A LUT can be represented as

$$\left\{\left(R_j, G_j, B_j, \hat{R}_j, \hat{G}_j, \hat{B}_j\right) \forall j \in [0, P^3]\right\}$$

with P being the resolution of the LUT, e.g. $P=64$. Each color value has a given fixed bit depth M, e.g. $M=10$ bits. A LUT for calibration of the digital chain in "film look" is calculated as follows:

$$\left(R_j, G_j, B_j\right) = \left(\frac{p\ \left(2^M - 1\right)}{P-1}, \frac{q\ \left(2^M - 1\right)}{P-1}, \frac{j\ \left(2^M - 1\right)}{P-1}\right)$$

$$\left(\hat{R}_j, \hat{G}_j, \hat{B}_j\right) = f_{RGB}^{TARGET}\left(f_{XYZ}^{REF}\left(R_j, G_j, B_j\right)\right)$$

with

$$p = j \bmod P^2;\ \ q = j \bmod P;\ \ 0 \le j < P^3$$

where f_{RGB}^{TARGET}

is the inverse device model of the target device and f_{XYZ}^{REF} the forward device model of the reference device.

Errors in the generation of the film look LUT can be a second source for errors in digital "film look". They may be caused by characterization errors (see above) and inherent linearization or quantization errors of the LUT.

Film Look Validation

This paper proposes a series of new subjective tests for validation of "film look" reproduction on digital displays in a DI postproduction workflow. The goal is to evaluate the precision of color reproduction on a target display (a digital projector) calibrated to "film look". The so-called Hypothetical Reference Circuit (HRC), i.e. the processing to be validated, consists in film scanning, LUT application and digital projection according to Figure 1.

Figure 2 shows how subjective film look evaluation is situated in the framework of film look generation. Subjective tests are applied to images displayed by (a) the film chain (digital images printed to film and then projected) and (b) by the digital chain (images projected by a digital projector) set to film look. The Film chain consists of film recording (printing a master negative from a Digital Intermediate) and film printing (copying a positive from the master negative). Negative and positive processes, printer parameters, film stocks have to be constant for all film used in the test. The digital chain contains a color box that applies a color transform to the RGB signal such that the target display is set to film look. The film look is generated by a 3D LUT that is calculated from a film chain device model and a digital chain device model such that the digital chain reproduces "film look". Additionally to subjective tests, "film look" can also be validated by objective measurements, for example using colored patches.

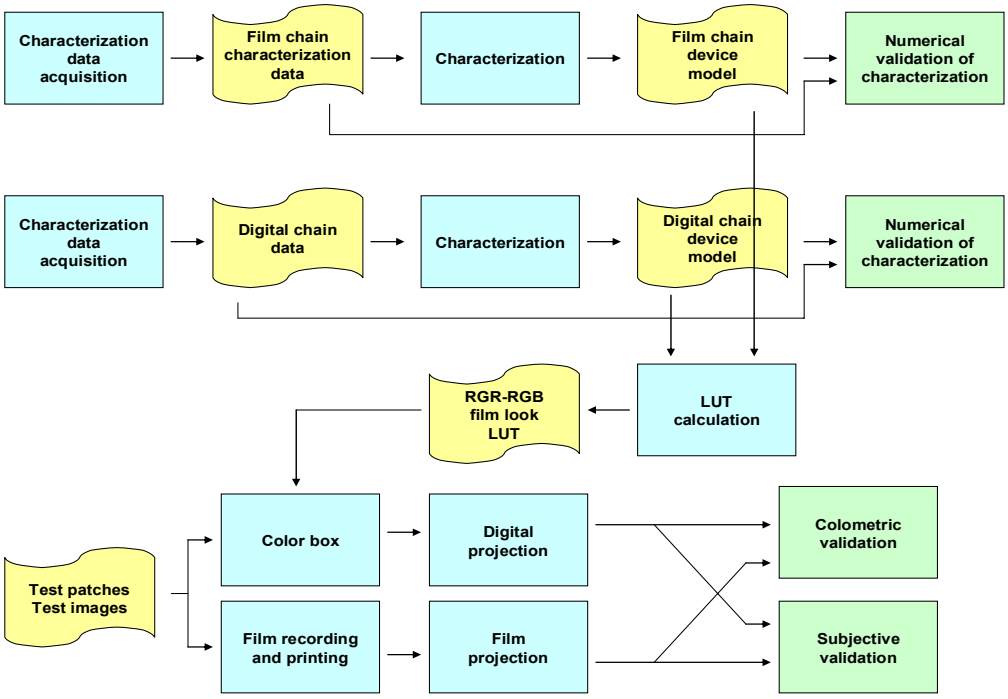

Figure 2. Validation and subjective evaluation of "film look" on digital display devices

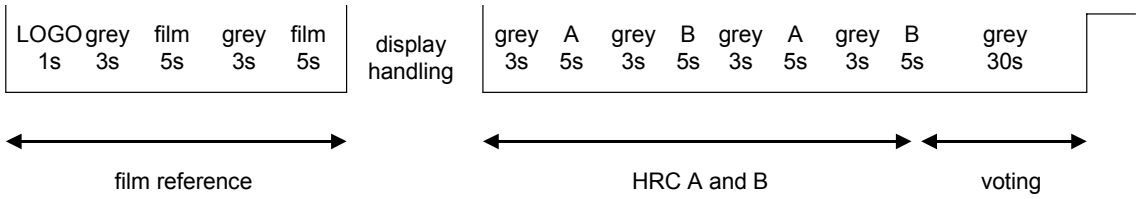

Figure 3. Basic test cell of test no.

Subjective psycho-visual test methods require human viewers, expert or non-expert, to rate the quality or difference in quality of two clips. In most testing scenarios these two clips differ in the fact that one will be the reference and the other will be processed in some manner. Subjective assessment can be a costly and time-consuming process, but one, however, that yields accurate results for any given evaluation. This type of assessment is particularly necessary in critical situations such as final product evaluation and standardization processes where quality must be assured. Subjective assessment methods have been used reliably in the past to evaluate video image quality.[2]

We propose a test methodology having the following characteristics:

1. **Use of real film content.** To enable final assessment comparable to real application case, short segments of real film content with high importance to color quality is used for the tests.
2. **Selection of content according to use cases.** The segments of test content will be selected corresponding to specific use cases such as "hue of skin color", "saturation of blue sky" and "tone of night scene" in order to cover all important regions of film gamut and allow for technical validation.
3. **Comparison of two stimuli to film.** Since viewers are biased when watching and recognizing the film reference version of test content (grain, jitter, weave), tests will include two digital stimuli after showing the film stimulus while assessments are done on the two digital versions with respect to film.
4. **Use of a small test persons group.** As compromise between statistical evaluation and limited availability of specialized test persons, the size of test group is limited to 12 persons.
5. **Restriction to US market**. For this first evaluation, content selection, criteria and experts will be chosen according to ethnical and cultural habits of Northern America.

Three tests are proposed:

No.	Number of stimuli	Order of stimuli	Type of stimuli	Scale type	Rel. ITU-R	Bias
1	3	tem-poral	moving picture	quality	DSCQS	Low
2	2	side-by-side	still picture	impair ment	DSIS, SDSCE	Yes
3	2	"	"	free text	-	?

Test 1 is designed to give a result of high confidence since observers are not biased (introduction of 2 unknown digital sources). Test 2 has a proven bias (film and digital sources are easily recognizable) but is close to the real application case of color correction. Test 3 has been added to collect natural language type comments of golden eyes.

Test No. 1: Moving Picture Test

This first test compares three versions of moving picture: film, digital film look under test and an existing digital film look reference; projects different content versions one after the other; asks for quality assessment according to a quality scale. We call this method Double-Stimulus Continuous Relative Quality Scale method (DSCRQS) which is derived from the Double-Stimulus Continuous Quality Scale method (DSCQS) proposed by Rec. ITU-R BT.500-10[3]

Figure 3 shows the basic test cell. A test cell begins with a test logo to indicate the start of a new test cell. A 10sec segment of test content is first projected as film two times and interlaced with 3sec medium gray sequences. Then projection is changed to digital media and the same segment of content is projected iteratively during 5sec for HRC A and B interlaced with 3sec medium gray sequences and followed by 30s medium grey for voting.

The test persons are asked to note A and B quality on a grading scale shown in Figure 4 according to specific questions on hue, saturation, contrast and white temperature. The viewers are allowed to assess during the voting period only. A test is run with 6 persons at once according to the test set-up shown in Figure 5.

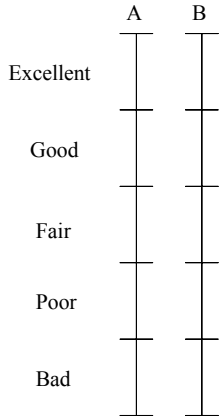

Figure 4. Quality grading scale for test no. 1

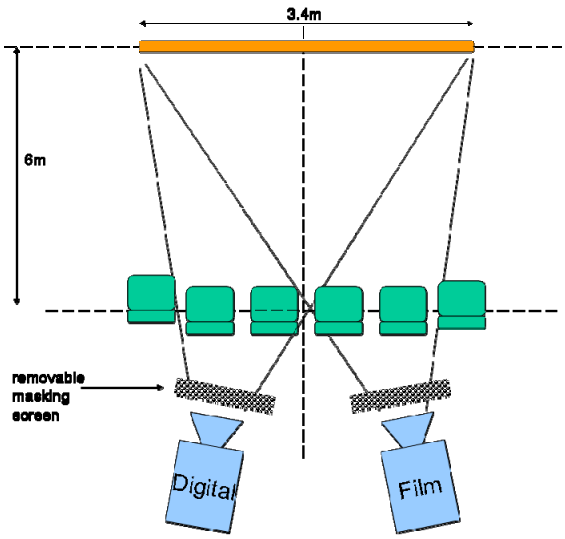

Figure 5. Viewing room set-up for subjective test no. 1 (derived from ITU-R BT.500-11

Test No. 2: Still Picture Butterfly Test

This second test compares two versions of a still picture: film and igital film look, projects different picture versions side-by-side (Rorschach butterfly,[4] asks for artifact assessment according to an impairment scale.

This test is derived from the Double-Stimulus Impairment Scale (DSIS) method and the Simultaneous Double Stimulus for Continuous Evaluation (SDSCE) method. From DSIS is taken the presentation of a reference version and a version under test. From SDSCE comes the side-by-side projection. Instead of video as in DSIS and SDSCE, still pictures are used to come close to the real application case of color correction. Figure 6 shows the basic test cell.

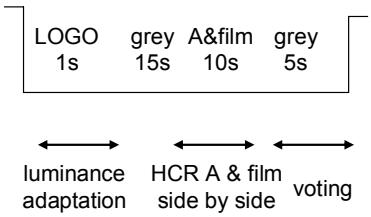

Figure 6. Basic test cell for subjective test no. 2

A test cell begins with a test logo to indicate the start of a new test cell. A 15sec medium grey period for adaptation is followed by the side-by-side presentation of the reference (right side: film) and HRC A (left side: digital film look). At the end, 5sec of medium grey are displayed for voting. The test persons are asked to assess the impairment of HRC A with respect to the film reference on a grading scale shown in Figure 7 according to specific questions corresponding to the use cases. A single test is run with 3-4 persons according to the test set-up shown in Figure 8. The viewers are allowed to assess during the voting period only.

0	The same
+1	Slightly different
+2	Different
+3	Much different

Figure 7. Impairment scale for subjective test no. 2 (modified from ITU-R BT.500-11)

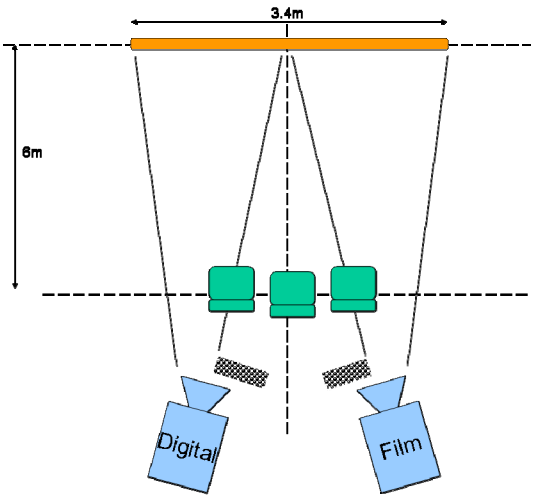

Figure 8. Viewing room set-up for subjective tests no. 2 and 3 (modified from ITU-R BT.500-11)

Test No. 3: Still Picture In-Depth Butterfly Test

This third test compares two versions of still pictures (film and digital film look), projects the different picture versions side-by-side, asks for in-depth assessment with unlimited time and free comments. Test no. 3 is very similar to test no. 2 except that presentation and voting time is unlimited and comments are in free text. This test responds to the fact that test persons are experts. Figure 9 shows the basic test cell. A test cell begins immediately with the image under test, the observers adapt to the image. The duration is unlimited, the presentation is side-by-side of the reference (right side: film) and HRC A (left side: digital film look). Presentation and assessment takes place at the same time. The test persons are asked to note comments in free text. A single test is run with 3-4 persons according to the test 2 set-ups shown in Figure 5.

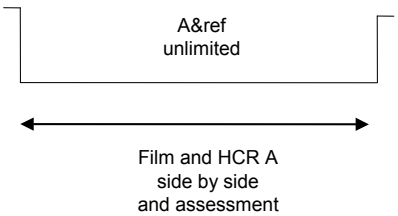

Figure 9. Basic test cell forsubjectivel test no. 3

Results

In a very first experiment, test no. 2 has been used to evaluate the effect of changing the film projector's bulb. After device characterization and LUT calculation, the bulb was replaced by a new one. The test was done after a 8h burn-in period. The change of projector's open gate white (projector without film) measured at beginning of the tests with respect to characterization time was $\Delta x=0.029$ and $\Delta y=0.002$ in CIE xy coordinates.

The scores of the test persons are in the interval from 0 ("the same") to 3 ("much different") according to Figure 7. Mean score for all 12 pictures is 2.05 with a standard deviation of 0.68. Before presenting a more in-depth analysis, some remarks about the limited validity of this test:

- As explained in above, test no. 2 derives from DSIS test and results are biased.
- The number of test persons is only 5 which is not sufficient for statistical analysis.
- One of the test persons was influenced (participation at a similar test one day before).

Figure 10 shows the scores of the 12 images, each score being the answer to a specific question. Scores are normalized to each test person's standard deviation.

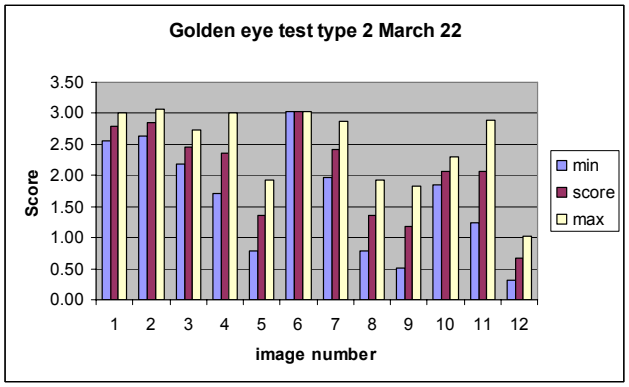

Figure 10. Mean scores and standard deviation intervals

When classifying the questions into four groups corresponding to four quality criteria, the mean scores shown in Table 1 for these criteria can be calculated.

Table 1: Mean Scores versus Quality Criteria

Criterion	Mean score	Standard variation
Contrast	1.79	0.96
Saturation	2.22	0.78
Hue	2.24	0.25
White temper.	1.17	-

The following results can be concluded:

1. The result with highest certainty (standard deviation 0.25) is on hue: it is assessed as "different" (mean score 2.24). The score on hue is worse than all other scores. The change of the bulb seems to influence mainly hue.
2. A score with less certainty (standard deviation 0.78) is the assessment of saturation: it is perceived in average to be "different" (mean score 2.22).
3. A score with even less certainty (standard deviation 0.96) is the assessment of contrast: it has a mean score of 1.79.
4. Only asked for one image, the white temperature is assessed to be "slightly different" (score of 1.17).

As mentioned, this test is biased and statistically not exploitable. Anyway, it shows that most impact of bulb change is on hue. The bad scores for saturation and contrast can be due either to indirect influence of hue changes on human perception and to the bias inherent in this test. The white temperature seems to be less affected. This can be explained by the fact that the tested scene has a white with high luminance, wheras hue deerivations were more observable at lower light levels.

Sample comments of test persons were:

- It should be explained to test persons that projected film was printed from the DI (as opposed to a copy from Original Negative).
- Butterfly presentation is not preferred by people involved in film process (opposed to people in DI process).
- Limited resolution of digital projection influences the assessment.
- Not matched contrast influences assessment of hue and saturation and vice versa.
- Overall hue of digital film look seemed to be warmer, including more yellow or less blue.

References

1. J. Stauder, L. Blondé, Introduction to cinematographic color management, IEE European Conference on Visual Media Production, CVMP-04, London, March 15-16, 2004.
2. Video Quality Expert Group (VQEP), www.vqeg.org
3. ITU-R BT.500-10, Methodology for the subjective assessment of the quality of television pictures, March 2000.
4. H. Ellenberger, The Life and Work of Hermann Rorschach (1884-1922), Bulletin of the Menninger Clinic, 18 (1954), 172-219; Reprinted many times; e.g. in Beyond the Unconscious. Essays of Henri F. Ellenberger in the History of Psychiatry (ed. M. Micale, Princeton UP 1993), 192-236.
5. A. Trémau, H. Konik, P. Colantoni; "Color Imaging Management in Film Processing", Internet Imaging V, part of SPIE conference Electronic Imaging, proceedings SPIE vol. 5304, January 19, 2004.

Author Biography

Jürgen Stauder received in 1999 the PhD degree from University of annover in the field of computer vision. He then stayed for 18 months with INRIA in Rennes, France, before joining Thomson R&D France. His research interests are computer vision, color science and computer raphics with application to video asset management, color management and compression.