Automatic emulation of human experts for estimating chromatic quality in digitization of graphical documents

Jesús Robledano-Arillo[a,*], Valentín Moreno-Pelayo[b], José Manuel Pereira-Uzal[c]

[a] Department of Library and Information Science, Universidad Carlos III de Madrid, C/ Madrid, 126, 28903 Getafe (Madrid), Spain. Phone: ++34 918561251. E-mail: jroble@bib.uc3m.es

[b] Computer Science and Engineering Department, Universidad Carlos III de Madrid, Avda. de la Universidad, 30, 28911 Leganés (Madrid), Spain. Phone: ++34 916249107. E-mail: vmpelayo@inf.uc3m.es

[c] Universidad Carlos III de Madrid, C/ Madrid, 126, 28903 Getafe (Madrid), Spain. Phone: ++34 918561251. E-mail: info@jpereira.net

Keywords

Document digitization

Photography

Image quality assessment

Machine learning

C4.5 algorithm

Visual algorithms

Abstract

This work aims to provide a critical examination of different approaches to creating models of automated quality control systems for digital images in digitization projects for photographic heritage collections. It investigates the feasibility of using machine-learning algorithms that work on sets of images previously evaluated by experts to obtain models on which to construct a high performance visual algorithm. We analyzed the data collected after conducting a psychometric experiment in which four human experts evaluated a set of three series of 300 degraded images by assigning each image to different quality classes. This analysis concludes that it is not possible to talk about commonly used simplistic models based on continuous acceptance ranges for colour metrics on an isolated basis, and therefore that it is necessary to

investigate more complex models. This study demonstrates that a model based on a machine learning rule-based system employing the CIE 1976 or CIEDE 2000 metrics along with the hue, saturation and lightness colour perceptual attributes emulates the human image quality experts with a high degree of efficacy, above 85%, opening an interesting way to get higher performance visual algorithms to automatically evaluate image quality in the context of digitization of photographic collections.

1. Introduction

In the context of heritage digitization of photographs and other documents with graphical value, a strict perspective of quality has taken hold that conceives of digital images as faithful representations at the physical and perceptual level: the images must faithfully represent the physical characteristics of the original physical documents and their appearance, under determined conditions of perception, during the digital capture process. Only this way can they be used for the functions of custody, conservation, reproduction, analysis, study and dissemination they are meant to support, within certain ethical criteria that do not approve of any change in the plastic characteristics or reinterpretation of the iconic and plastic messages (Martínez & Muñoz, 2002; Ruiz, 2004; Robledano, 2011a, 2011b). This strict perspective has important implications when it comes to proposing a procedure for controlling the quality of the digitizations, as its application introduces the need to operate on two planes: a physical plane and a perceptual one.

In terms of the first, the level of quality can be measured objectively in a simple way by applying certain physical attributes to the image that have been widely studied in recent decades in the fields of imaging engineering and colour science and technology (resolution, colour coding error, dynamic range, OECF, etc.), as well as by measuring the impact on the digital signal of a series of distortions that can affect the performance of the attributes (noise, chromatic aberrations, geometric distortions, compression artefacts, etc.). There have even been various attempts to systematize these characteristics in the context of digitization of cultural collections (Frey & Reilly, 1999, 2006; FADGI-Still Image Working Group, 2010). Based on a set of pre-selected physical attributes, it is possible to construct a multidimensional quality model that enables computation of the quality of the digital image of an original document digitized together with one or more reference cards with respect to its corresponding physical original. Different multidimensional models have been used to compute quality based on the measurements obtained for attributes, such as the Generalized Weighted Mean Hypothesis or the Minkowski metrics (Engeldrum, 1995). As shown in equation 1, quality (Q) in this type of model can be approached as a function that calculates the Euclidean distance of the degraded image (x) with respect to an ideal image (y) in an n-dimensional space, with the dimensions being the attributes (i) included in the tests, weighted through its weighting coefficients (p).

$$Q(x,y) = \sqrt{\sum_{i=1}^{n} \left((x_i - y_i).p_i\right)^2}$$

(1)

On the perceptual plane, image quality is the possibility of generating visualizations or reproductions based on this image that evoke in the user an overall perception similar to that which he or she would experience observing the original document under certain determined and controlled observation conditions, and without any type of distortion. Overall appreciation of quality at the perceptual level is a subjective process that commonly occurs when a human observer regards the physical document alongside a reproduction or visualization of its corresponding digital image. The human observer will attempt to quantify the degree to which the digital image departs perceptually from its corresponding original under certain normalized viewing conditions according to standards (ISO, 2008, 2009b). The introduction of a human evaluator is very costly, creating the need in many massive digitization projects to devise automated quality control systems that can replace the expert human observer in the quality comparison phase but do not reduce the high performance of an experienced human observer in terms of evaluating the perceptual proximity between the original and its corresponding digital image.

Given the ease of computing attributes and distortions of a physical nature, one important line of research on how to develop these systems has been the attempt to connect the physical and perceptual performance levels, such that the overall quality of an image at the perceptive level can be automatically derived through the use of easily computable physical measures by working with a limited number of attributes and ranges of values, and with highly efficient processes. The term visual algorithm is commonly used to refer to this type of mathematical model. The problem stems from the fact that it is not easy to derive perceptual fidelity from physical fidelity directly. Many efforts to create a robust visual algorithm at the level of subjective human perception of overall quality based on physical attributes have failed because they did not sufficiently consider the multiplicity of elements and complex interrelationships that underlie this phenomenon in a sufficiently exhaustive quality model (Engeldrum, 2004; Zhou, Bovik, & Ligang, 2002). The performance of these types of approaches suffers for diverse reasons, such as the non-linearity of human perception of quality problems, the use of attributes lacking a strong degree of correspondence with the perceptual appreciation of quality (Engeldrum, 2004), handling the attributes independently without considering that they are mutually interactive (Lee, 2005), or not incorporating the influence of a series of subjective factors that condition visual interpretation of the image and which have been widely studied (Fairchild, 2004).

The application of multidimensional scaling methods that enable analysis of the complex interactions underlying the quality attributes of the images has been explained by Lee (2005), who references some that have obtained physical descriptors for psychophysical attributes (Martens, 2002; Pellacini, Ferwerda, & Greenberg, 2000). Also the application of machine-learning methods through which one can infer determinant quality attributes and their

interrelationship models for the automation of image quality control systems, but in areas outside the context of activity that is our focus and employing experimental graphical databases, such as LIVE or TID2008, whose characteristics differ from those of the type of heritage object we address in our research. Machine learning is an attractive approach for image classification based on quality assessment. Various methods have been proposed within what tends to be called Machine Learning-based Image Quality Measure, such as that described by Charrier, Lézoray, and Lebrun (2012), and Narwaria, Lin and Cetin (2012). These contributions have limited reach for the heritage objectives we propose in this work.

In the context of graphical heritage objects, automated digitization quality systems have been based essentially on tests that can be classified as belonging to the physical level, exclusively using a limited set of attributes of this type for which certain previously determined value acceptance ranges are established. If we consider the main papers published on this question in the field of documentary heritage, treatment of the problem of the connection between the physical and perceptual levels seems to have been addressed in an overly simple way, in most cases leaving aside the subjective perceptive component inherent in the quality control process for a graphical medium. Many of the papers have focused on identifying and proposing metrics for exclusively physical attributes, but without going into depth about a perceptual model for overall image quality that would provide guidance when establishing the systematic acceptance ranges for the performance of these attributes and their complex interrelationships during the act of perception (Williams, 2000, 2002, 2003, 2010; Puglia, Puglia, Reed, & Rhodes, 2004; Still Image Working Group, 2010; FADGI-Still Image Working Group, 2010; Bureau Metamorfoze, 2007; Dormolen, 2010; Nationaal Archief, 2010).

Our paper focuses on an attempt to establish a valid working line for automated creation of highly efficient visual algorithms that can be used in quality control systems for digital images from the digitization of works of a graphical nature, and which will make it possible to overcome the limitations of systems based on multidimensional models that use a predefined set of quality attributes together with their ranges of acceptance values. Due to the breadth of this objective, we will address only the use of colour attributes. We attempt to demonstrate that it is possible to model the perceptual value judgments of an expert, or set of expert human evaluators, as regards the perceptual proximity in colour between a digital image and its corresponding physical original, through an efficiently computable visual algorithm based on the combined use of standardized colour metrics and perceptual colour attributes. The complexity of the interactions between colour attributes that occur in the perceptive act makes automation of the process of obtaining the visual algorithm necessary. To do this, we propose applying a machine-learning method based on the induction of rules that do not require pre-definition of the most determinant quality attributes and their acceptance range beforehand, and which can work on a data set obtained from the real processes of human evaluation to be modelled. For this paper, we have applied the machine-learning algorithm to the data obtained from an experimental set of images previously evaluated by a pre-selected group of human experts in image evaluation. The use of an experimental set was costlier, but it allowed us to configure the test to obtain more accurate knowledge about the suitability of the colour attributes and metrics selected for building a robust visual algorithm for this type of evaluation.

2. Methodology

2.1. Phase I. Test of evaluation with human experts

The test consisted of emulating a real quality evaluation process with human experts, applying certain ideal evaluation context conditions, according to the standards for carrying out quality evaluation, through comparison of the physical originals to the corresponding onscreen digital images: ISO 3664 (ISO, 2008), 12646 (ISO, 2009) and 20462-3:2012 (ISO, 2012).

Three photographic images on paper that were representative of the type of documents found in many photography collections were used: modern positive photographic materials in colour with glossy and matte finishes, and old photographic materials hand-coloured in ink. We selected diverse iconic motifs on the premise that the image motif influences the perception of quality. We chose human figures and landscapes with typical icons (sky, clouds, grass, forest, water). With this difference in representational motifs, we can attempt to analyze how the difference in motif influences the judgment of quality. In the table we present the codes assigned to each image and its iconic description.

448	**449**	**550**
Matte colour photographic print. Mountain and water landscape. Late twentieth century.	Monochromatic photographic print. Hand-coloured. Human Portrait. Early twentieth century.	Glossy colour photographic print. Human Portrait. Early twenty-first century.

Table I. Description of experimental images.

We then created the digitized masters of the original images directly using a digital single-lens reflex camera and applying colour management through customized ICC colour profiles in order to obtain images with high fidelity in colour and contrast at the colorimetric and densitometric level. Likewise we took spectral samples of small controlled areas of the surfaces of the images by applying a template that allowed exact identification of the sampled area and reflected-light spectrophotometry with the intention of also using these areas, along with the control card patches, to measure the physical and perceptual colour attributes to be used for subsequent derivation of the visual algorithm.

Based on the masters, a series of between 303 and 300 degraded images per physical original were created by editing their HSL perceptual values: Hue (H), Saturation (S) and Lightness (L). This created a degradation sequence that contemplated a sufficiently broad scale of perceptible changes in these three colour-description variables. To do this, the images were converted to the HSL colour space and progressively degraded in these three variables, from 20 to 19 for hue (on a scale ranging from -100 to +100), from -39 to +39 for saturation (on a scale ranging from -100 to +100) or from -20 to 20 for lightness (on a scale ranging from -100 to 99). Likewise repeated images were generated so that we could measure the degree of consistency in the evaluations, analyzing how the evaluators' selective criteria varied over the course of the test, if applicable, and to be able to determine the probability of the evaluators

giving random responses during the test. The repeated images were distributed throughout each series of images to be evaluated.

The data for the images to be evaluated were recorded automatically by applying different comparison metrics for colour and image difference between the original photographs and the digital masters and their degradations, of which we selected only two for the tests we present in this paper: CIEDE 2000 – CIE00 – (Luo, Cui, & Rigg, 2001) and CIE 1976 L*a*b* colour-difference formula – CIE76 – (ISO, 2007). For calculating the colour differences between the digitized physical images and the digital images from the series of degraded ones, we used all the patches from the Colorchecker card and the three colour samples taken directly from the photographs. We will use MCIE76 and MCIE00 to refer to CIE76 and CIE00 metrics applied to the latter samples.

The visual evaluation of the experts was carried out based on the perception of the digital images reproduced on the monitor, making it necessary to minutely control all the elements that made up the viewing flow, which, in addition to the digital image, included the following: the calibration and ICC profile of the monitor; the conversion from the colour space of the image to the colour space of the monitor by the operating system's colour management system (CMS); the quality of the monitor and of the conditions of its viewing environment; the quality of the booth for viewing the physical originals and its viewing conditions. The screen interface was designed using the Adobe Bridge program so that only the image being evaluated appeared, with a narrow band along the left edge showing the images in the group to allow the evaluators to select the next image to view and navigate through the batch. The quality value applied and the coded name of the image were displayed underneath. The Colorchecker card used to create the masters and the original itself were placed in the booth in a position that was very similar to that in the test images. The intensity of the grey background colour of the screen was made to coincide with that of the booth.

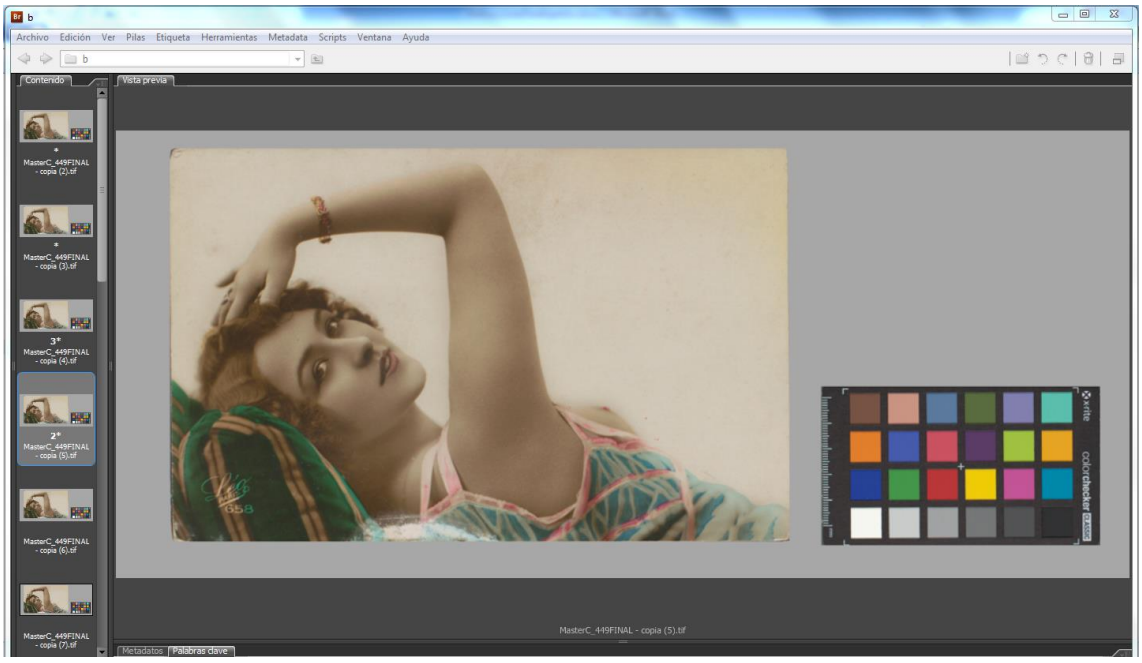


Figure 1. Evaluation interface showing image 449.

Based on the quality detected, the experts were able to assign a score to each image based on a scale with three values: 1 (the image would not pass a professional quality control measuring the proximity in the appearance of colour and contrast between an image on the screen and an image on paper); 2 (the image would pass the quality control but with a less rigorous criterion); and 3 (the image would pass the quality control with a rigorous criterion). In the interest of simplifying our first analytical approach in this study, we consolidated values 2 and 3 so that we would be working with only two quality classes: valid image and invalid image.

Four experts were selected who met the condition of being professionals with extensive experience in the sectors of professional photography and graphic arts (8, 14, 15 and 16 years of work experience in evaluation). Part of their daily work consists of visually evaluating the quality of images and comparing the proximity between the onscreen image and the printed copy. The team of experts was given a sufficient period of instruction to understand the type of quality evaluation required in the field of documentary heritage.

2.2. Phase II. Data analysis

With the collected data, two types of analysis were carried out:

1) Analysis of consistency in the quality judgments of each evaluator.

There were two objectives: to detect and estimate, percentagewise, errors due to lack of consistency in the evaluations of the human experts participating in the test, and to be able to compare the error percentages of the experts with those of the rules-based system we obtained afterward through machine learning. We applied two parameters that allowed us to measure the degree of consistency in the evaluations of each human evaluator (intra-evaluator) and between evaluators (inter-evaluator).

a) Intra-evaluator consistency error.

This type of consistency error is indicative of the application of random evaluation processes at some point in the test or of changes in the quality criteria employed in the course of this. The factors that can cause both behaviours are multiple and vary widely in type: fatigue prior to or during the course of the test, lack of concentration, lack of interest, lack of engagement, etc. An evaluator, even an experienced one, can also adjust his or her criteria diachronically during the evaluation process in response to the errors being detected or the order in which images with varying degrees of distortion are presented. Consistency was measured using repeated images inserted into the series. This was calculated, as shown in equation 2, by adding up, for each expert, all the consistency errors that occurred over the course of the three series of images and the total number of repeated images to discover the percentage represented by the first compared to the total number of images. A consistency error is understood to mean a difference in the scores assigned to identical repeated images.

$$C = 100\ x\ \frac{\sum_{i=1}^{n} e(a_i) \neq e(b_i).1}{n} \quad (2)$$

where n is the number of repeated images, $\mathcal{I}$ the set of images, $e: \mathcal{I} \rightarrow \mathbb{N}$ the assessment function, and $a_i\ and\ b_i$ the identical representations of the repeated image i

b) Inter-evaluator consistency.

Lack of consistency at this level is due mainly to the use of different criteria or degrees of strictness during evaluation. We applied three indicators:

- Experts' degree of strictness. This was calculated by discovering the percentage of images selected as valid out of the total number of images evaluated.
- Degree of consistency between experts with respect to the coincidence of scores for the same images. To calculate this, we measured the percentage of coincidence between each pair of experts in the three degrees of scoring permitted in the test. This meant measuring, when an expert assigned a certain value to the images in the series, what proportion of the other experts coincided. For example, of the images to which expert 1 assigned value 2, in what percentage did expert 2 also assign value 2. We should point out that the percentage does not necessarily have to coincide in reverse: continuing with our example, expert 2 could also have assigned value 2 to many other images to which expert 1 assigned other values.
- Degree of coincidence in the scores for all the images by the four evaluators. Represented by the sum of images where all the experts coincided on the same value and of the images where there was no such coincidence.

2) Regularity analysis in the behaviour of the four HSL colour perceptual values and of the CIE colour difference metrics in the evaluators' quality judgments.

We tried to detect whether or not regular intra- and inter-evaluator patterns existed in the dispersion of the values for the different attributes that would explain the quality criterion applied by the experts, and which colour perceptual attributes would best enable modelling of the experts' behaviour. The existence of these patterns would facilitate the work of obtaining visual algorithms, based on which highly efficient evaluation systems could be generated that would approximate human evaluation processes in terms of accuracy. After analyzing the results, we will have to be able to determine whether it is feasible to generalize quality models based on fixed acceptance ranges for the colour difference metrics and perceptual attributes considered in this study. This aspect is highly relevant, as many quality control systems currently being applied are based on this model.

To do this, we analyzed the ranges of acceptance values (quality score 1) and rejection values (quality score 2 or 3) in the metrics and HSL variables for each of the experts and images,

attempting to detect some regularity in them. Subsequently we comparatively analyzed the behaviour of the values for the ones degraded in the HSL variables with respect to the values of the colour difference metrics in the valid and invalid image groups.

2.3. Phase III. Application of a machine-learning method for obtaining and validating a visual algorithm

The detection of regular behaviour patterns in the analyzed variables reinforced the idea of the utility of applying machine-learning techniques for obtaining a visual algorithm that models their behaviour patterns with a high predictive capacity. The emulation of the human responses to image quality assessment has been approached through systems based on artificial neural networks with some success (Tchan, Thompson & Manning, 1999), but we have chosen to apply a rule induction algorithm because the information that the rules give us will be highly relevant for understanding the visual algorithm, determining its computation efficiency, and gaining knowledge about the act of quality perception by a human expert. We have also considered the greater flexibility provided by a rule-based approach to change or add parameters (Zhiqing & Yang, 1999). We applied a rule induction method, the C4.5 machine-learning algorithm (Quinlan, 1993), using Weka machine-learning software (Witten & Frank, 2005). This algorithm derives an organized rule-based system in the form of a decision tree that is easy to read and comprehend.

We used all the data from the three images and those from two of the experts: expert 1, as the most inconsistent, and expert 4, as the most consistent. Values 2 and 3 remained consolidated in a single class to allow us to work with a binary-type attribute. The number of positive instances (valid images) and negative ones (invalid images) was compensated by 50% to avoid polarization of the model towards the most numerous class. The compensation was done by repeating the data records for the positively evaluated images. We used only five variables: the expert score, the CIE76 metric and the three HSL colour perceptual attributes. The method applied to validate the resulting rule was crossed validation with 10 folders.

3. Results and discussion

3.1. Response consistency of the experts

a) Intra-evaluator consistency error.

Expert 1	**Expert 2**	**Expert 3**	**Expert 4**
20%	15.22%	15.22%	10.87%

Table II. Intra-evaluator consistency error.

b) Inter-evaluator consistency.

- Experts' degree of strictness.

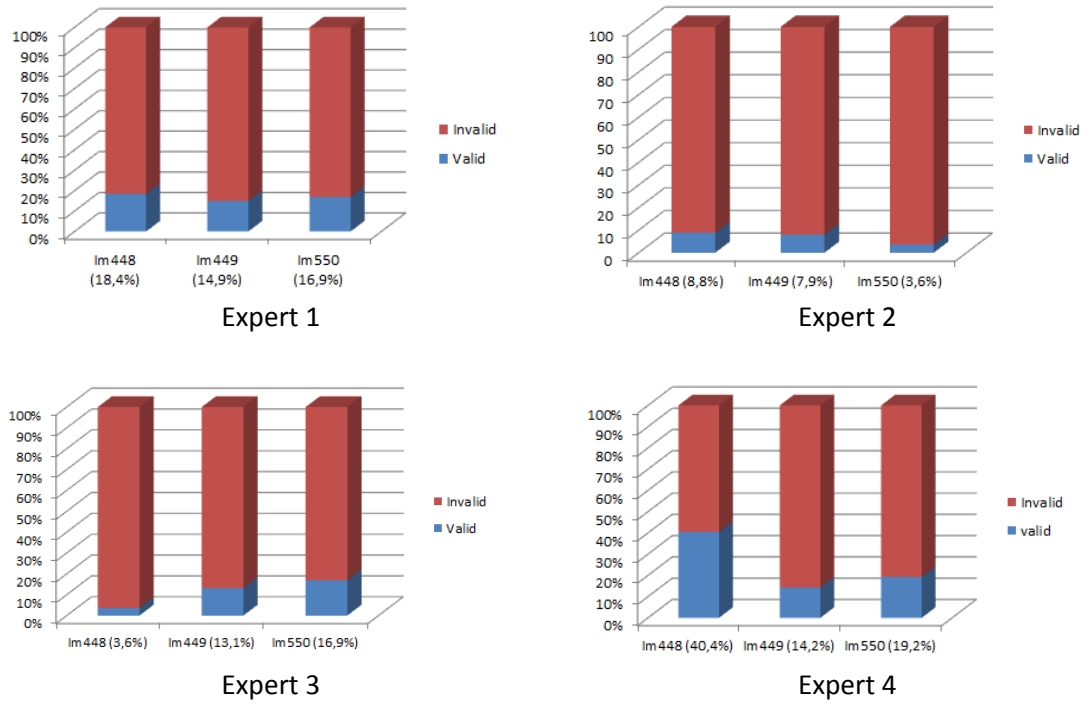


Figure 2. Experts' degree of strictness, represented by the percentage of images considered valid for each of the four experts.

- Degree of consistency between experts with respect to the coincidence of scores for the same images.

In tables III to VI, we present the average of the percentages obtained for the three images.

	Average % agreement in assigning value 1.	Average % agreement in assigning value 2.	Average % agreement in assigning value 3.
Expert 1 and 2	95.97	14	0
Expert 1 and 3	91.03	25.75	9.5
Expert 1 and 4	82.87	38.6	25.57

Table III. Average of the agreement percentages from expert 1 compared to the other experts.

	Average % agreement in assigning value 1.	Average % agreement in assigning value 2.	Average % agreement in assigning value 3.
Expert 2 and 1	85.7	31.33	0

Expert 2 and 3	88.96667	21.8	0
Expert 2 and 4	78.63333	43.56667	0

Table IV. Average of the agreement percentages from expert 2 compared to the other experts.

	Average % agreement in assigning value 1.	Average % agreement in assigning value 2.	Average % agreement in assigning value 3.
Expert 3 and 1	87.03	36.77	3.17
Expert 3 and 2	95.5	13.97	0
Expert 3 and 4	82.66	51.7	39.67

Table V. Average of the agreement percentages from expert 3 compared to the other experts.

	Average % agreement in assigning value 1.	Average % agreement in assigning value 2.	Average % agreement in assigning value 3.
Expert 4 and 1	91.73	34.9	14.3
Expert 4 and 2	97.43	18.63	0
Expert 4 and 3	95.5	32.266	13.27

Table VI. Average of the agreement percentages from expert 4 compared to the other experts.

- Degree of coincidence in the scores for all the images by the four evaluators.

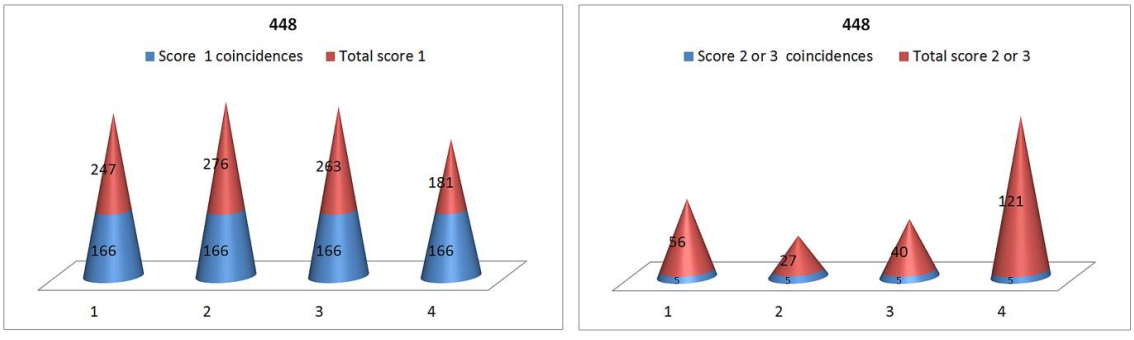


Figure 3. Scoring coincidences for the images in the 448 series by all the experts.

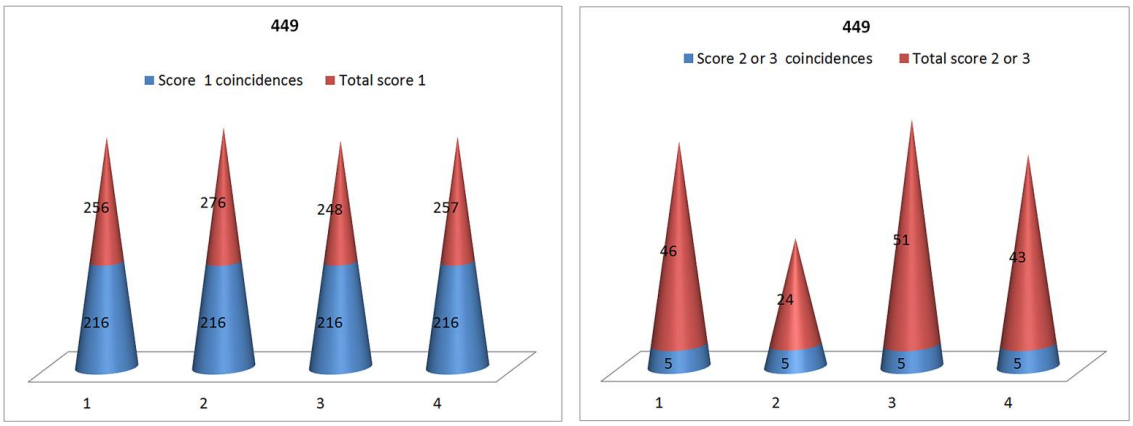


Figure 4. Scoring coincidences for the images in the 449 series by all the experts.

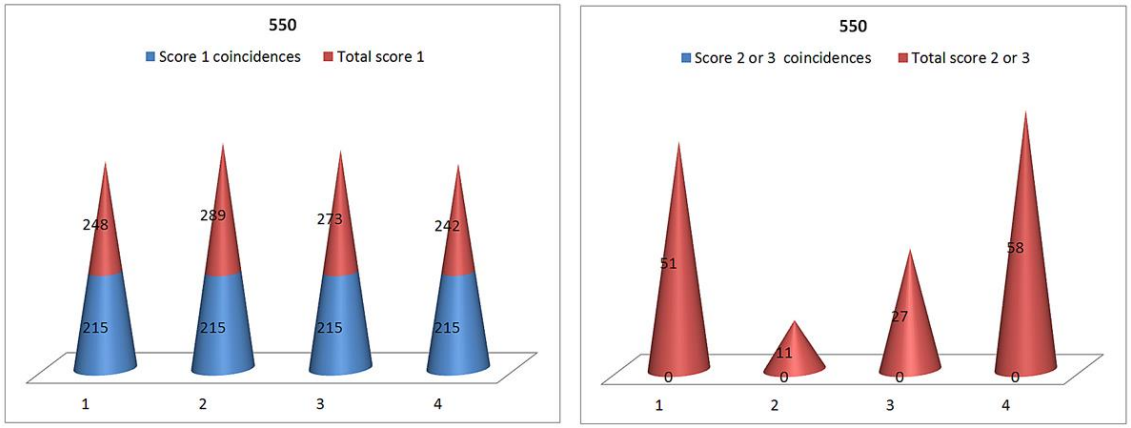


Figure 5. Scoring coincidences for the images in the 550 series by all the experts.

There were intra-evaluator consistency errors with all the experts, with a variability of between 5 and 10%. The error percentages are not exceptionally high, and therefore we can rule out a sustained random response and assume that the experts performed their evaluations guided by their perception of the quality of the images and applying a quality criterion. The expert with the most experience in evaluating documentary heritage, number 4, was the most consistent, suggesting that baseline training was a factor in the performance in this parameter.

It seems that the degree of strictness of the four experts was generally very high, as the percentages, except for image 448, were under 20%. There only appears to have been a disparity of opinions for image 448, where the difference between the percentages was high, and for one of the experts (expert 2), the strictest, for image 550.

In all the series of images, we can appreciate how the degrees of coincidence are much higher for the images considered invalid, which is explained by the fact that there was a much higher number of images considered invalid in all the series and for all the experts. We can say that the coincidence is in general low, and therefore we cannot talk about uniformity in the criterion of the four experts. As a result, without a period for the participating human experts

to agree on a common criterion beforehand, a quality control will always yield a low consistency rate, providing little reliability and consistency.

The degrees of intra- and inter-evaluator consistency are indicative of the difficulty of achieving very high efficiency percentages with a visual algorithm that models their behaviour in a very accurate way, as the algorithm will also model the inconsistencies. There were fewer inconsistencies at the intra-evaluator level, and therefore achieving high performance rates by obtaining individual algorithms for each expert is feasible. In a real case, it would be necessary to try to analyze why the inconsistencies occur, improving the training of the experts to increase the consistency levels before proceeding to obtain the algorithm.

3.2. Regularities in the behaviour of the HSL colour perceptual values and the CIE colour difference metrics in the judgments of the evaluators

First we present a table with the threshold values for all the metrics and colour-perceptual variables for each image and, following this, these same data referenced to each expert. These data help calibrate the scope of the acceptance ranges for the images.

	Value	**CIE76**	**CIE00**	**MCIE76**	**MCIE00**	**Hue**	**Satur.**	**Light.**
448	Max.	16.51	9.72	15.38	13.49	19	39	19
	Min.	0.97	0.66	3.87	3.25	-20	-39	-20
449	Max.	16.55	9.35	14.21	11.47	19	39	19
	Min.	1.11	0.80	3.09	2.45	-20	-39	-20
550	Max.	16.79	9.62	14.31	11.78	19	39	19
	Min.	0.88	0.69	5.10	4.33	-20	-39	-20

Table VII. Maximum and minimum values of the above metrics in the three sets of images.

	Value	**CIE76**	**CIE00**	**MCIE76**	**MCIE00**	**Hue**	**Satur.**	**Light.**
448	2 and 3 (58 images)	0.97 a 11.68	0.66 a 7.71	3.76 a 9.08	4.57 a 10.84	- 4 a 10	-31 a 8	-12 a 18
	1 (245 images)	1.04 a 16.51	0.68 a 9.72	3.87 a 15.38	3.25 a 13.49	-20 a 19	-39 a 39	-20 a 19
449	2 and 3 (46 images)	1.18 a 10.86	0.90 a 7.84	3.09 a 8.18	2.45 a 5.79	-5 a 9	-25 a 14	-3 a 19
	1 (254 images)	1.11 a 16.55	0.80 a 9.35	3.21 a 14.21	2.63 a 11.47	-20 a 19	-39 a 39	-20 a 18
550	2 and 3 (51 images)	0.88 a 10.74	0.69 a 7.84	5.22 a 11.77	4.45 a 9.40	-6 a 6	-23 a 8	-7 a 19

	1 (248 images)	0.88 a 16.79	0.69 a 9.62	5.10 a 14.31	4.33 a 11.78	-20 a 19	-39 a 39	-20 a 17

Table VIII. Value ranges for the expert 1.

	Value	CIE76	CIE00	MCIE76	MCIE00	Hue	Satur.	Light.
448	2 and 3 (27 images)	1.09 a 11.63	0.74 a 6.97	3.87 a 8.92	3.25 a 7.41	-2 a 11	-13 a 14	-7 a 13
	1 (276 images)	0.97 a 16.51	0.66 a 9.72	3.98 a 15.38	3.35 a 13.49	-20 a 19	-39 a 39	-20 a 19
449	2 and 3 (24 images)	1.11 a 7.58	0.80 a 5.91	3.30 a 7.52	2.45 a 6.31	0 a 4	-17 a 5	-7 a 13
	1 (images)	1.18 a 16.55	0.89 a 9.35	3.09 a 14.21	2.63 a 11.47	-20 a 19	-39 a 39	-20 a 19
550	2 and 3 (11 images)	0.88 a 8.47	0.69 a 6.70	5.48 a 10.34	4.67 a 8.38	0 a 6	-10 a 2	-3 a 13
	1 (289 images)	0.88 a 16.79	0.69 a 9.62	5.10 a 14.31	4.33 a 11.78	-20 a 19	-39 a 39	-20 a 19

Table IX. Value ranges for the expert 2.

	Value	CIE76	CIE00	MCIE76	MCIE00	Hue	Satur.	Light.
448	2 and 3 (40 images)	0.97 a 10.74	0.66 a 7.79	3.98 a 9.08	3.35 a 7.52	-2 a 8	-10 a 14	-7 a 19
	1 (263 images)	1.08 a 16.51	0.68 a 9.72	3.87 a 15.38	3.25 a 13.49	-20 a 19	-39 a 39	-20 a 16
449	2 and 3 (30 images)	1.18 a 10.86	0.90 a 7.84	3.21 a 8. 18	2.63 a 5.79	-4 a 3	-13 a 9	-2 a 19
	1 (248 images)	1.18 a 16.55	0.89 a 9.35	3.09 a 14.21	2.45 a 11.89	-20 a 19	-39 a 39	-20 a 18
550	2 and 3 (27 images)	0.88 a 7.39	0.69 a 5.85	5.22 a 9.50	4.45 a 7.70	-1 a 2	-10 a 1	-11 a 13
	1 (273 images)	0.88 a 16.79	0.69 a 9.62	5.10 a 14.31	4.33 a 11.78	-20 a 19	-39 a 39	-20 a 19

Table X. Value ranges for the expert 3.

	Value	CIE76	CIE00	MCIE76	MCIE00	Hue	Satur.	Light.
448	2 and 3 (122 images)	0.97 a 13.88	0.66 a 8.32	3.87 a 12.12	3.25 a 10.41	-4 a 13	-35 a 14	-15 a 19
	1 (181 images)	1.08 a 16.51	0.68 a 9.72	4.3 a 15.38	3.60 a 13.49	-20 a 19	-39 a 39	-20 a 13
449	2 and 3 (43 images)	1.11 a 10.86	0.80 a 7.84	3.21 a 8.18	2.63 a 5.97	-4 a 6	-17 a 19	-7 a 19
	1 (257 images)	1.18 a 16.55	0.89 a 9.35	3.09 a 14.21	2.45 a 11.47	-20 a 19	-39 a 39	-20 a 18
550	2 and 3 (58 images)	0.88 a 10.221	0.69 a 7.42	4.45 a 9.13	5.22 a 11.42	-7 a 6	-17 a 2	-14 a 18
	1 (242 images)	0.88 a 16.79	0.69 a 9.62	5.10 a 14.31	4.33 a 11.78	-20 a 19	-39 a 39	-20 a 19

Table XI. Value ranges for the expert 4.

In the interest of reducing the amount of data to be viewed, we simplified the CIE metrics based on their correlation, selecting only one out of the most correlated metrics. We applied the Pearson correlation coefficient because of its suitability for the type of linear correlations we find between all the studied variables. In the following tables, we show the correlations between the CIE metrics used in the study.

	CIE00 y CIE76	MCIE00 y MCIE76	MCIE00 y CIE00	MCIE76 Y CIE76
448	0.93	0.99	0.72	0.57
449	0.92	0.97	0.68	0.52
550	0.93	0.96	0.89	0.87

Table XII. Pearson correlation results between CIE Delta E metrics from the Colorchecker patches and from the physical document samples.

All the correlations are significant, and therefore we chose to use the CIE76 metric, one of the most used in colour quality evaluation. In the interest of simplifying viewing, we used only image 448, considering that in light of the data for all the images provided in the previous tables, it is possible to generalize these conclusions for the three images in the study.

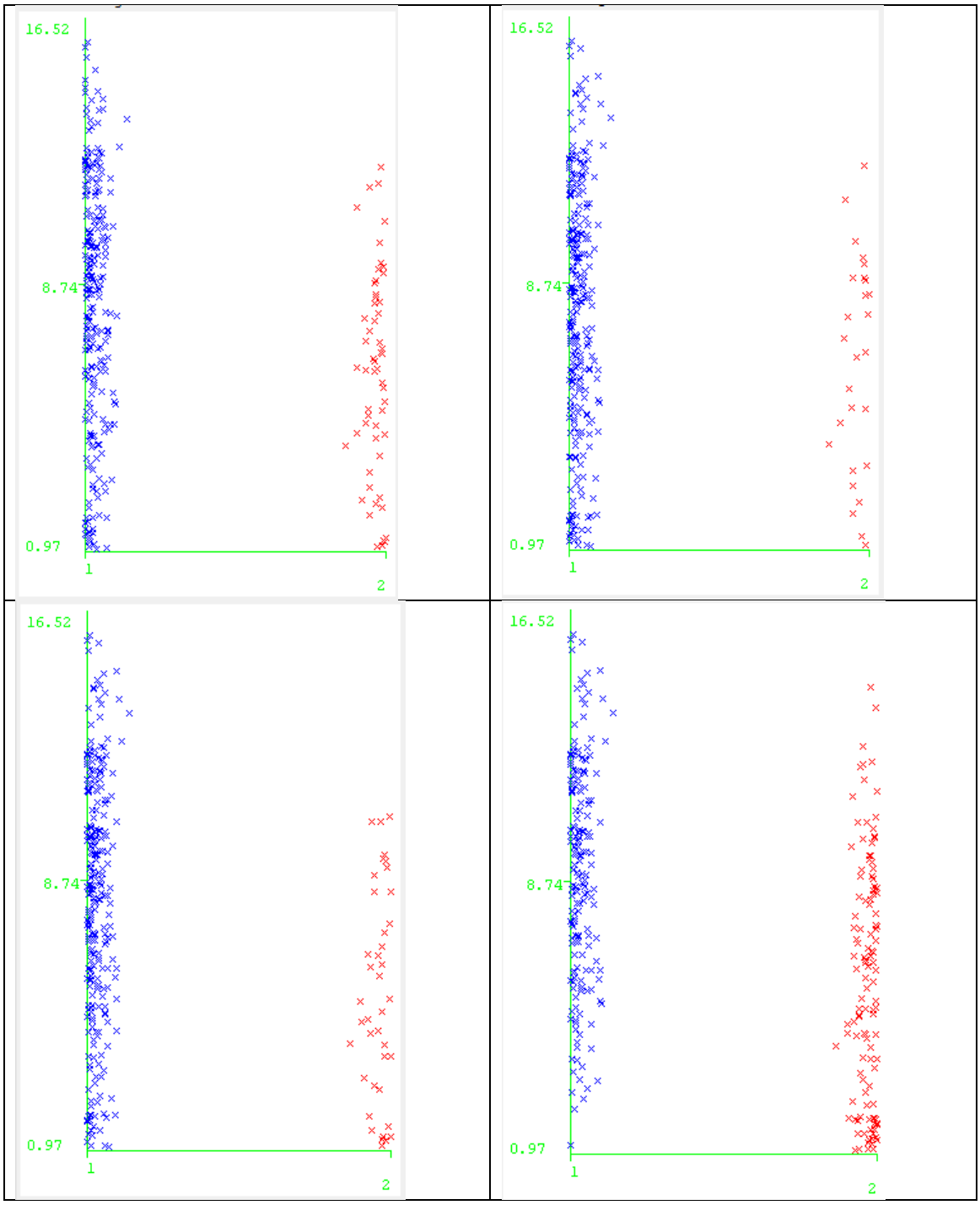


Figure 6. Distribution of valid (red) and invalid (blue) images in the CIE76 value ranges for the four experts (in order from 1 to 4). Image 448.

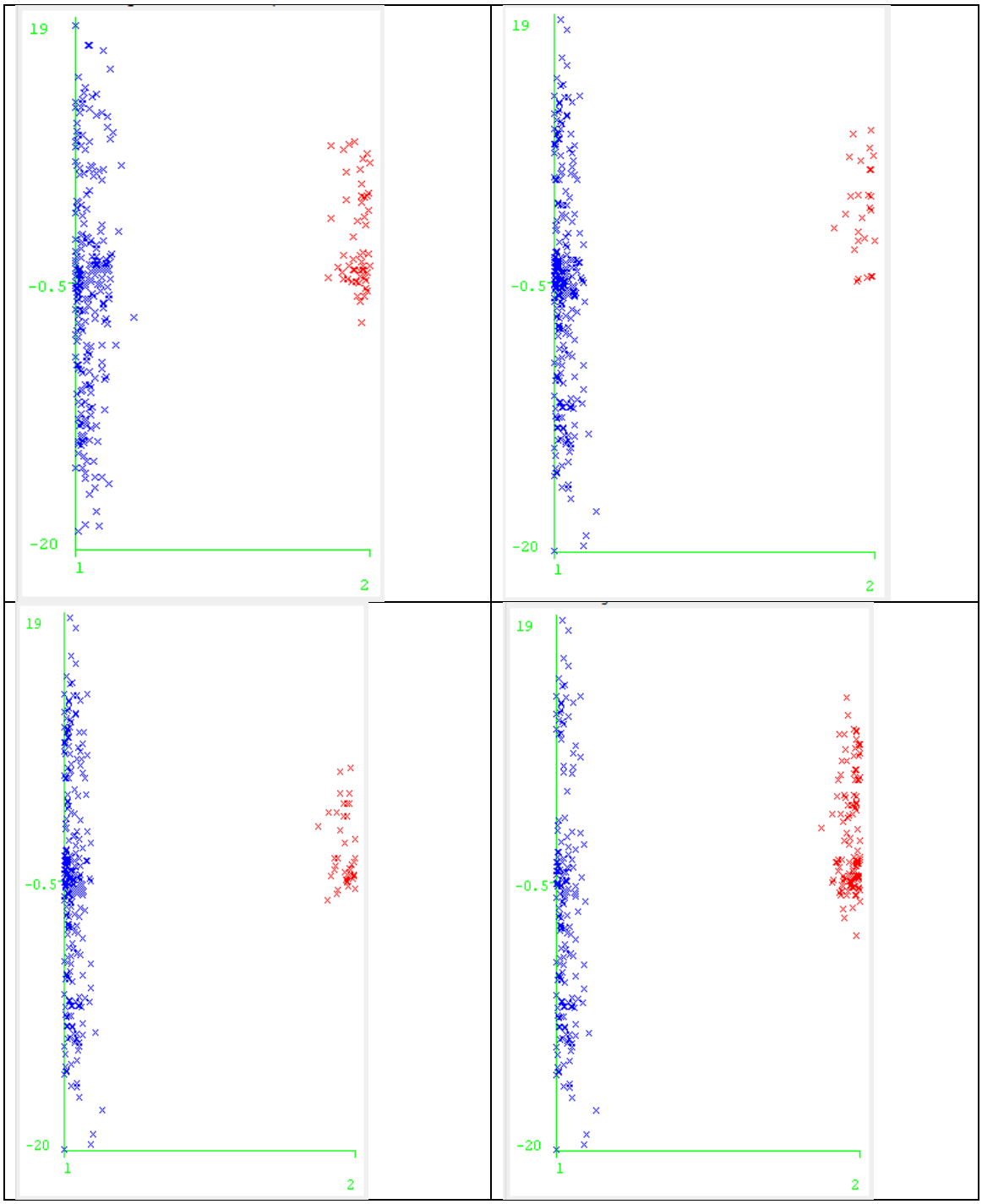


Figure 7. Distribution of valid (red) and invalid (blue) images in the hue value ranges for the four experts (in order from 1 to 4). Image 448.

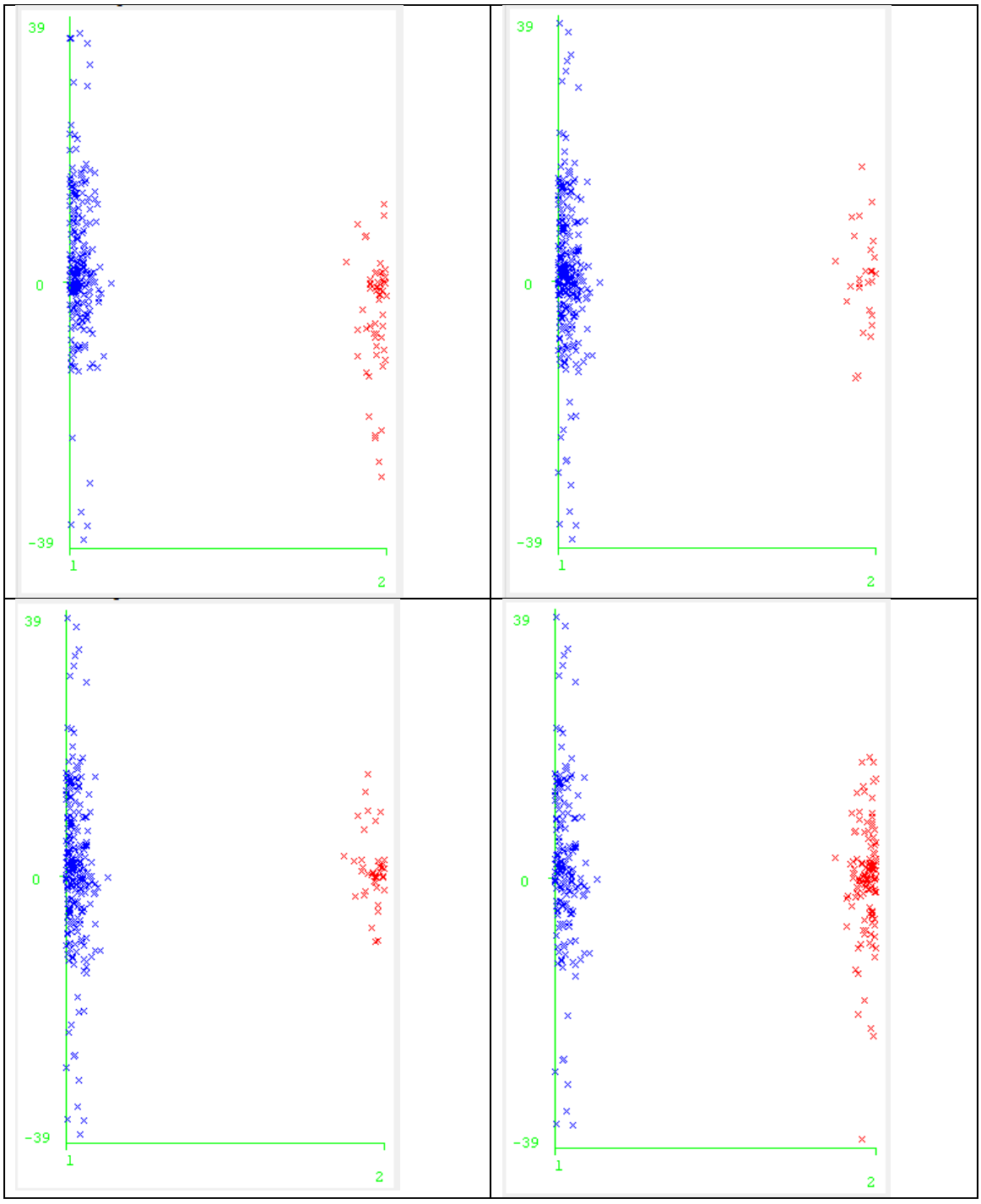


Figure 8. Distribution of valid (red) and invalid (blue) images in the saturation value ranges for the four experts (in order from 1 to 4). Image 448.

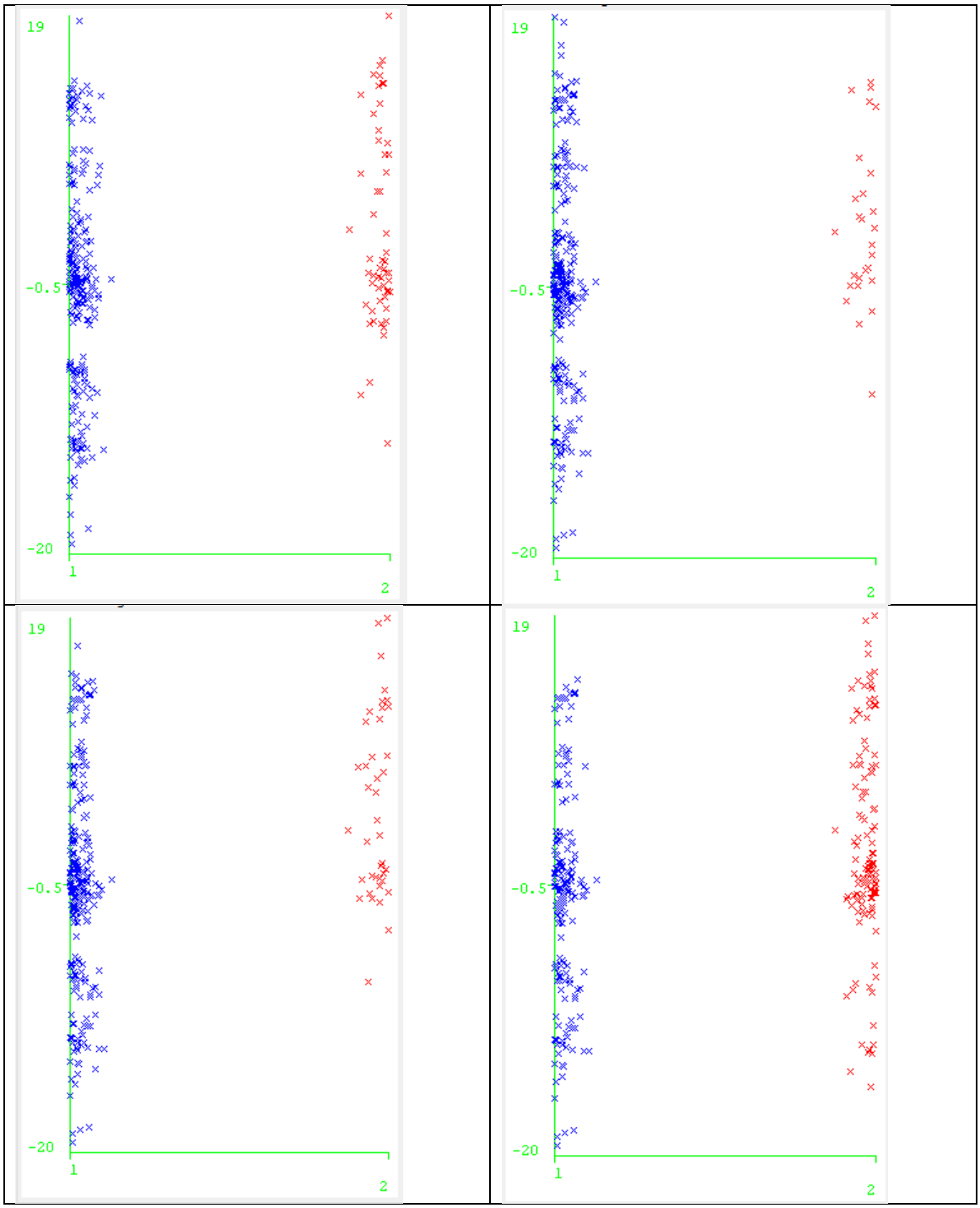


Figure 9. Distribution of valid (red) and invalid (blue) images in the lightness value ranges for the four experts (in order from 1 to 4). Image 448.

The pattern is very similar for the four experts and for the three images, except in the case of expert 4, who was less strict and accepted a wider range for the valid images. In all the metrics, except for the low end in CIE76, the range for invalid images includes the range for the valid ones, except for some internal discontinuity. The overlap in the ranges is very high, which

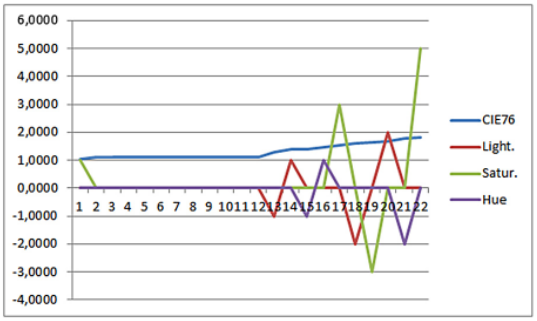
prevents the establishment of fixed acceptance ranges for a single metric or colour perceptual value in isolation. In any case, it would be possible to set a threshold value above which the image could be considered rejected, although it would not be possible to say anything about validity below this value. In this study, we have seen how this threshold value varies depending on the expert and on the image within the scores of a single expert. We compared to see whether this threshold value came close to being the same for the four experts by comparing the data from the tables and graphs, but this was not the case; the threshold value oscillated in the CIE76 metric between 10.7 and 11.6 for expert 1, between 8.47 and 11.6 for expert 2, between 7.39 and 10.8 for expert 3 and between 10.22 and 13.88, for expert 4. In the CIE00 metric, this was between 7.71 and 7.84 for expert 1, between 5.91 and 6.7 for expert 2, between 5.85 and 7.79 for expert 3, and between 7.42 and 8.32 for expert 4. It seems that the experts varied their criterion for each type of image, with the acceptance range for the diverse metrics not always having same value for the three images. Therefore, we can conclude that the iconic motif of the image is determinant with respect to the degree of strictness applied by the expert and the perception of colour and hue problems.

In view of the results, it would be necessary to review the utility of the fixed acceptance ranges in the CIE76 and CIE00 metrics found in many image quality control systems, as the acceptance ranges are much wider than those commonly considered in heritage quality control systems and admit a high percentage of invalid images; to discard these, it would be necessary to also consider performance in colour perceptual attributes and their interrelationships, aspects that do not seem sufficiently modelled in the CIE metrics we used.

In order to determine the degree of similarity in the correlation patterns between quality judgments and value variations in the parameters analyzed by the different experts, we studied what happens in the zones of overlap closely. The zones of overlap are the intervals within the values for a variable where as many valid images as invalid images were found. The purpose of this analysis was to determine the factors that cause an image within these zones to be considered valid or invalid by each expert, and whether a regular pattern exists in the behaviour of these factors that will help us obtain a model. For example, we have the case of accepted and rejected images in the CIE76 range between 0 and 4. We want to know if this is because there is a high degree of randomness in the quality evaluation when the degradation is not very evident, or because factors influencing the perception of quality related to the HSL colour perceptual variables exist.

We are going to analyze the role played by variability of HSL, so that the images are considered valid or invalid within the same interval, by considering the CIE76 metric and one of the images, 448. We did not use CIE00 because its results were practically the same as for the previous one, because the zones of overlap between CIE00 and CIE76 coincided 92.74% of the time and due to the high degree of correlation between the series of values for the two metrics. To do this, we analyzed the individual data for each expert and each image, dividing the values for the CIE76 metric into intervals, between ranges 1 and 8, and studying how the variation in the HSL variables behaves. The x-axis shows the image's order number, and the y-axis shows the value of the HSL and CIE76 variables. The images were sorted in ascending order by their CIE76 value. To simplify the results, we show only the data for the comparison

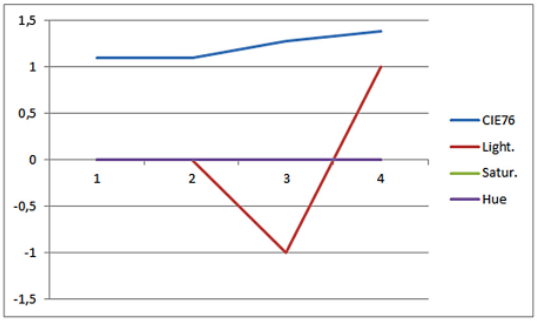
between experts 4 and 1 with image 448. We present the graphs of the valid and invalid images side by side to facilitate viewing of the regularities in the patterns for the two types.



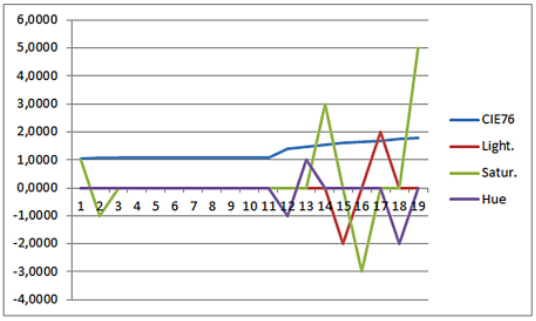
Expert 4. Delta 1. Valid images

None

Expert 4. Delta 1. Invalid images

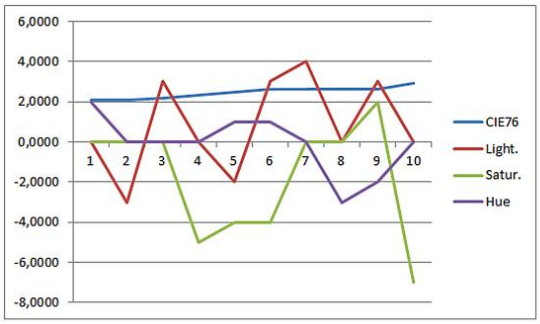


Expert 1. Delta 1. Valid images

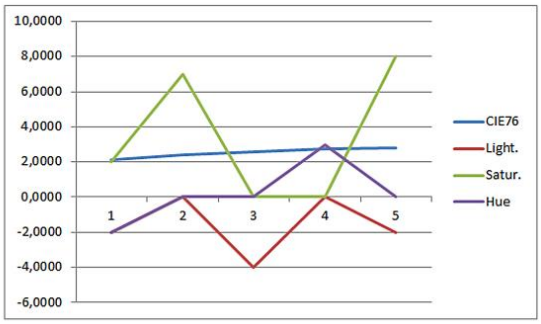


Expert 1. Delta 1. Invalid images

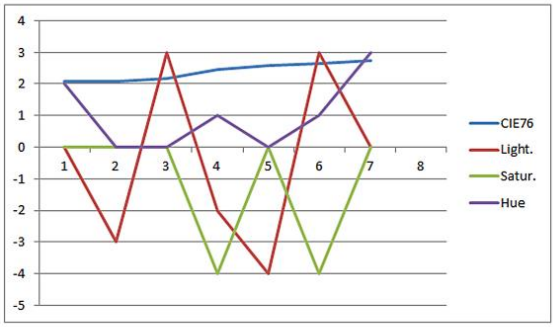
Figure 10. Delta 1 ranges.



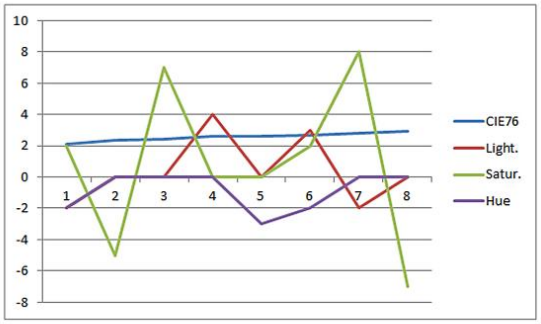
Expert 4. Delta 2. Valid images



Expert 4. Delta 2. Invalid images



Expert 1. Delta 2. Valid images



Expert 1. Delta 2. Invalid images

Figure 11. Delta 2 ranges.

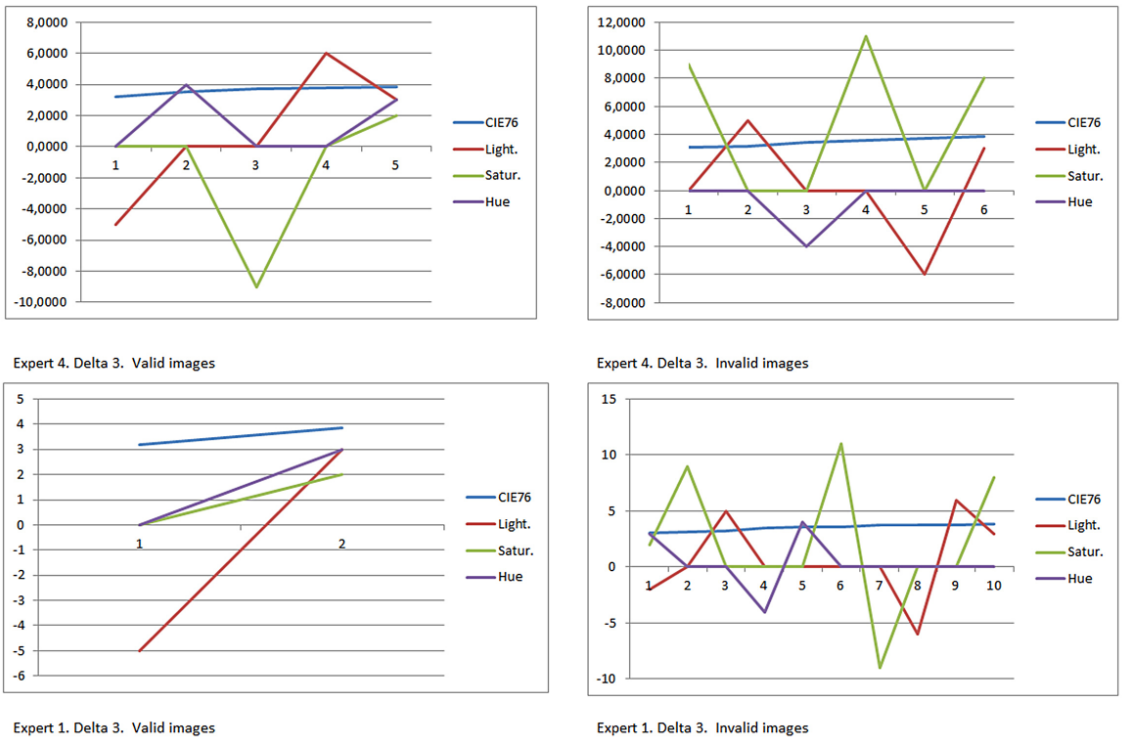


Figure 12. Delta 3 ranges.

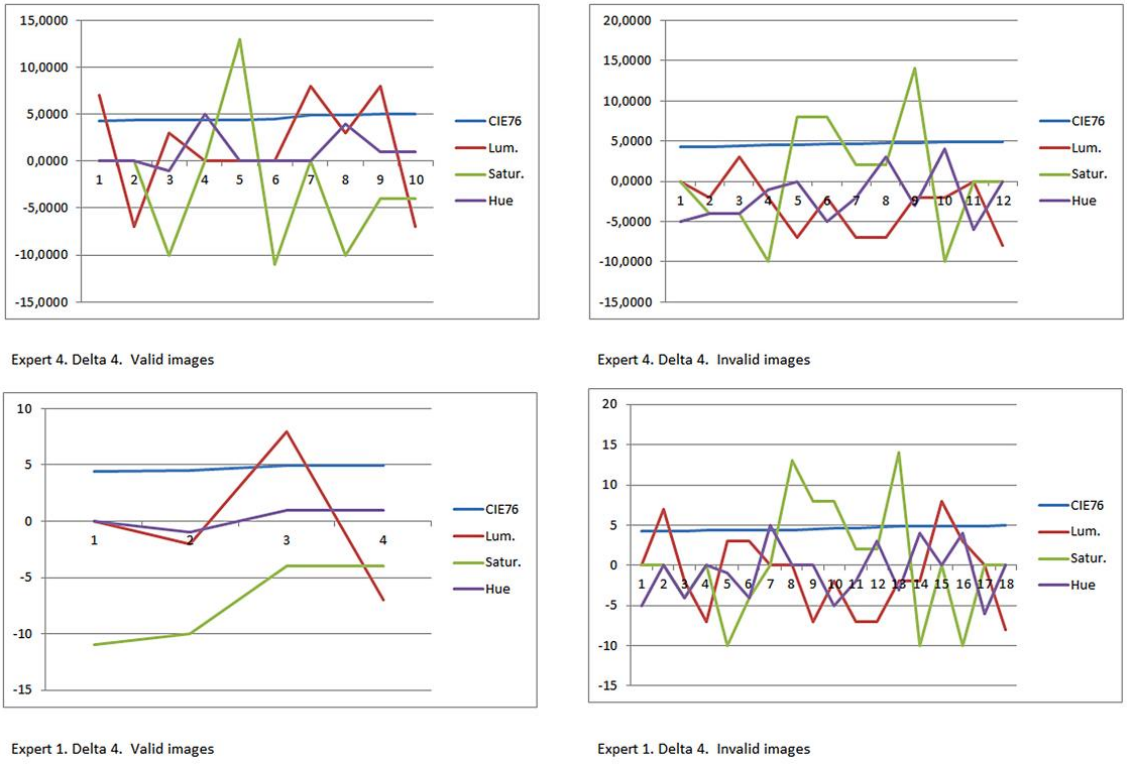
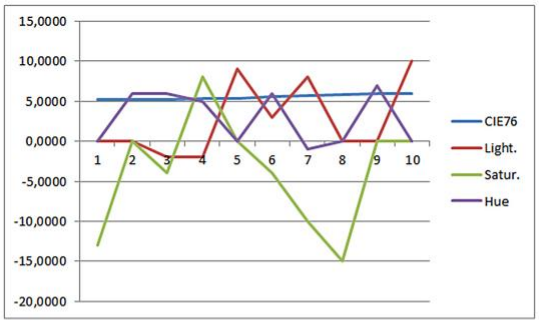
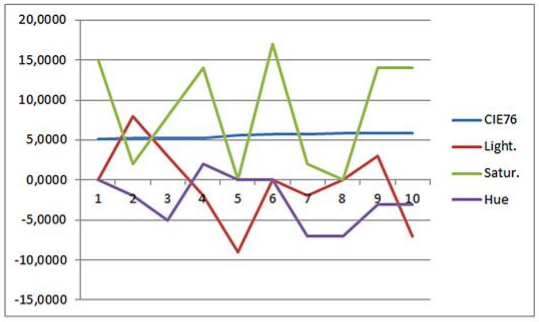


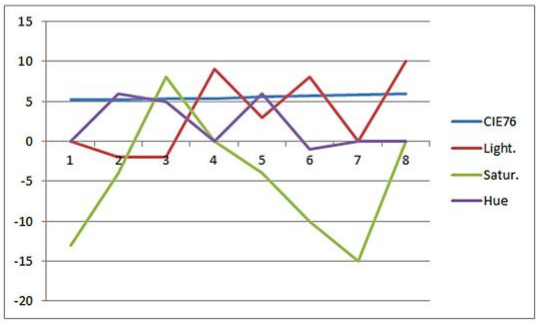
Figure 13. Delta 4 ranges.



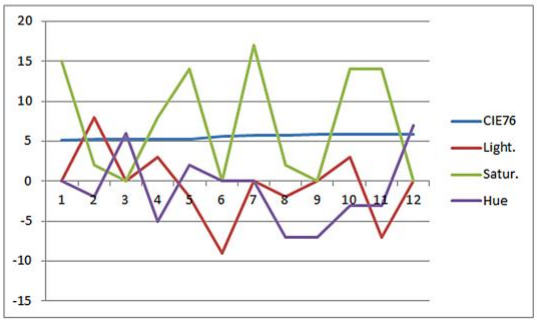
Expert 4. Delta 5. Valid images



Expert 4. Delta 5. Invalid images

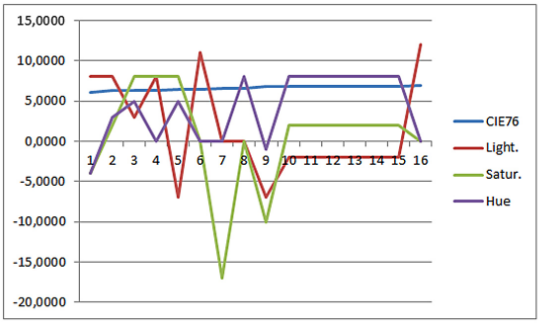


Expert 1. Delta 5. Valid images

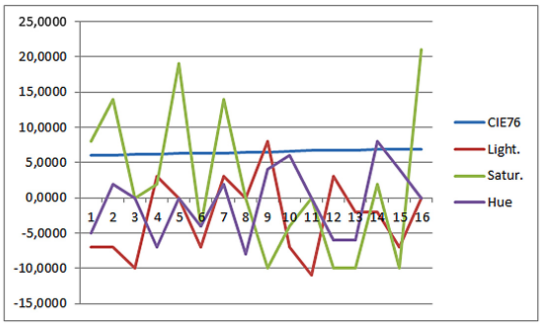


Expert 1. Delta 5. Invalid images

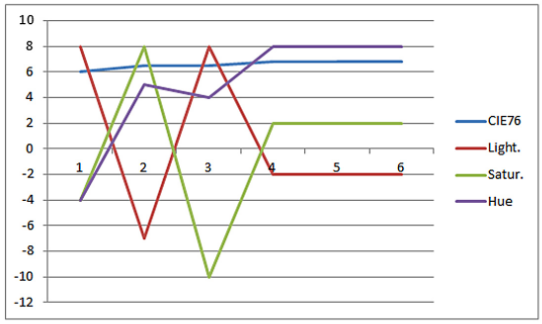
Figure 14. Delta 5 ranges.



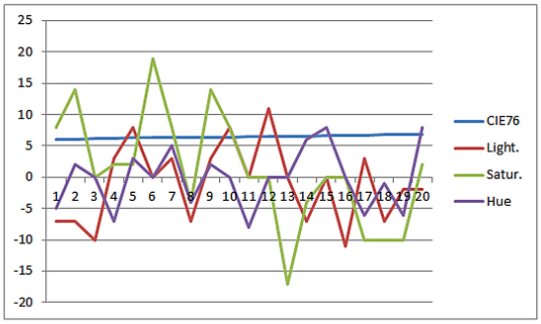
Expert 4. Delta 6. Valid images



Expert 4. Delta 6. Invalid images

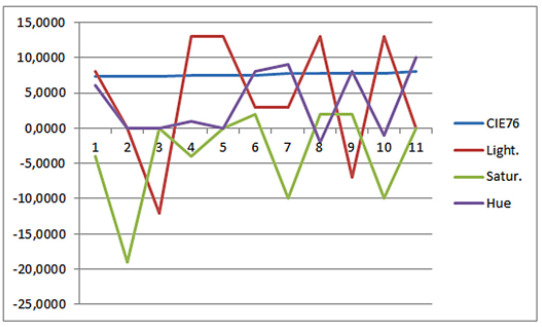


Expert 1. Delta 6. Valid images

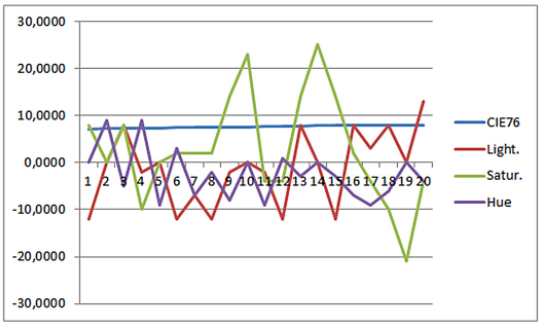


Expert 1. Delta 6. Invalid images

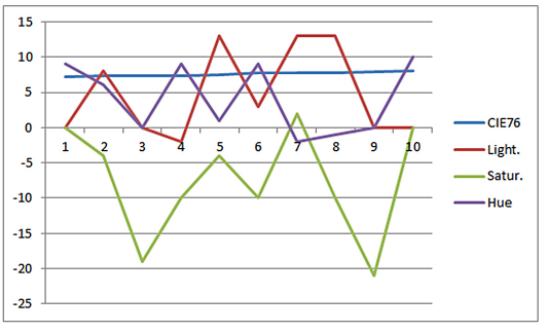
Figure 15. Delta 6 ranges.



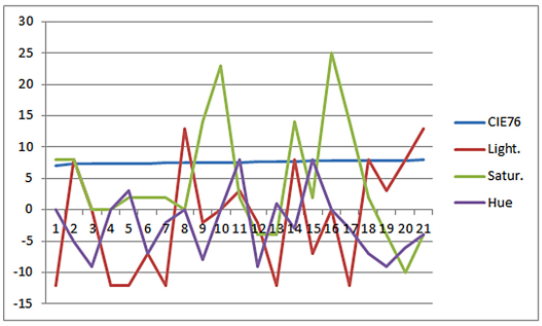
Expert 4. Delta 7. Valid images



Expert 4. Delta 7. Invalid images

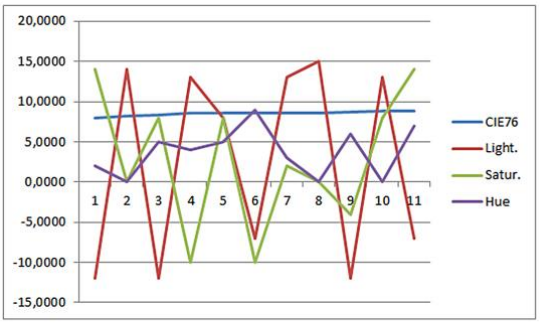


Expert 1. Delta 7. Valid images

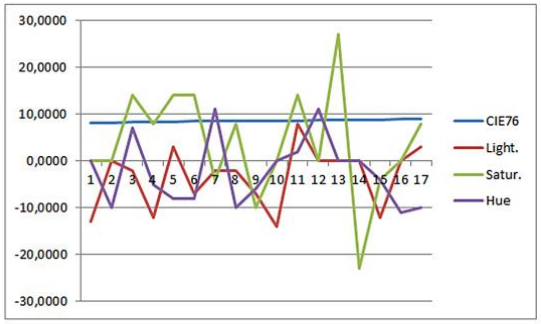


Expert 1. Delta 7. Invalid images

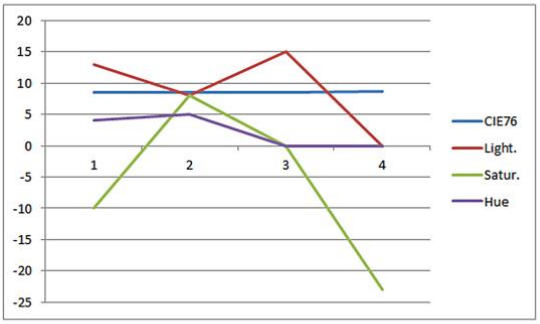
Figure 16. Delta 7 ranges.



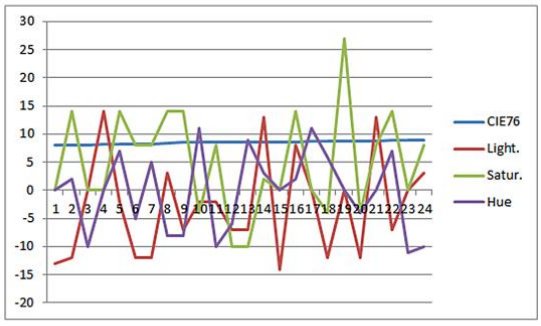
Expert 4. Delta 8. Valid images



Expert 4. Delta 8. Invalid images



Expert 1. Delta 8. Valid images



Expert 1. Delta 8. Invalid images

Figure 17. Delta 8 ranges.

If we look at these graphs, we can see how the numerical patterns of the HSL colour perceptual variables are very different within the same delta range between the images accepted and rejected by the two evaluators; although some coincidences also appear, there are very few. The coincidences represent anomalous behaviour, as they assume that the expert is applying the same conformity criterion to the valid images as to the invalid ones, and vice versa, but they are explainable if we consider the data for consistency in the response of the experts we included previously. If we assume the same inconsistency percentage as we had in the repeated images for the rest of the images, it is logical that we find repeated patterns of rejection within the acceptance patterns, and vice versa. But the differences found in the patterns reinforces the idea that it is not possible to base quality control models on fixed acceptance ranges for the CIE Delta E 1976 or CIEDE 2000 metrics without also considering the behaviour in the hue, saturation and lightness variables. This consideration is very important in the overlap ranges. Therefore, rigid models based on ranges of metrics considered in isolation cannot be used to obtain effective quality evaluation.

It is possible to confirm the existence of a similar numerical model in the HSL values for the valid and invalid images in experts 1 and 4, which becomes practically identical in the invalid ones as the delta increases. This progressive similarity is explainable because, there being a greater number of invalid images in the highest deltas, it is more likely that the coincidence between the two experts will increase gradually for the images they selected as valid and invalid. In delta 7, for example, the coincidence between expert 1 and expert 4 is 23 images out of a total of 32 images. If we return to the data shown above on the level of consistency between the experts, we can see how experts 1 and 4 are very consistent with each other, over 90% of the time for the invalid images and close to 40% for the valid images.

3.3. Performance of a machine-learning method for obtaining and validating a visual algorithm

Using algorithm C4.5, we obtained a set of rules that makes it possible to classify new examples of images as valid or invalid in the same way as the human expert whose evaluation data was used to infer the rules. To measure the degree of efficacy and efficiency, we used different indicators. These include the precision and recall rates. The first expresses the proportion among the images recovered by the rule-based system within a class, for example class 1 (invalid), of ones that are correct because they correspond to their class and ones that are not. The recall rate expresses the proportion of images of one class that were correctly assigned by the rule-based system compared to all the images corresponding to this class.

For the image 448:

Expert 1									
Total images	Correctly classified		Incorrectly classified		Precision		Recall		Number of rules
471	Total	%	Total	%	Valid	Invalid	Valid	Invalid	33
	406	86.383	64	13.617	0.784	0.984	0.987	0.752	
Expert 4									
Total images	Correctly classified		Incorrectly classified		Precision		Recall		Number of rules

728	Total	%	Total	%	Valid	Invalid	Valid	Invalid	40
	670	92.033	58	7.967	0.899	0.944	0.948	0.892	

Table XIII. Rule system performance for the image 448.

For the image 449:

Expert 1									
Total images	Correctly classified		Incorrectly classified		Precision		Recall		Number of rules
525	Total	%	Total	%	Valid	Invalid	Valid	Invalid	14
	482	91.8095	43	8.1905	0.898	0.942	0.948	0.886	
Expert 4									
Total images	Correctly classified		Incorrectly classified		Precision		Recall		Number of rules
515	Total	%	Total	%	Valid	Invalid	Valid	Invalid	13
	472	91.6505	43	8.3495	0.857	1	1	0.833	

Table XIV. Rule system performance for the image 449.

For the image 550:

Expert 1									
Total images	Correctly classified		Incorrectly classified		Precision		Recall		Number of rules
504	Total	%	Total	%	Valid	Invalid	Valid	Invalid	16
	470	93.4394	33	6.5606	0.885	1	1	0.867	
Expert 4									
Total images	Correctly classified		Incorrectly classified		Precision		Recall		Number of rules
474	Total	%	Total	%	Valid	Invalid	Valid	Invalid	19
	436	91.9831	38	8.0169	0.867	0.986	0.987	0.855	

Table XV. Rule system performance for the image 550.

In figure 18 we show the set of rules obtained automatically for the case with the highest success rate in correctness, image 550 and expert 1, where we can see how the ranges for the CIE76 metric are influenced in all cases by the range of values in the HSL variables.

```
CIE76 <= 8.7
|   Saturation <= 3
|   |   Hue <= -2
|   |   |   Lightness <= 5: 1 (30.0)
|   |   |   Lightness > 5
|   |   |   |   CIE76 <= 6.77: 1 (2.0)
|   |   |   |   CIE76 > 6.77: 2 (16.0/1.0)
|   |   Hue > -2
|   |   |   Hue <= 5
|   |   |   |   Lightness <= -3
|   |   |   |   |   Saturation <= -3: 2 (16.0/1.0)
|   |   |   |   |   Saturation > -3: 1 (14.0)
|   |   |   |   Lightness > -3: 2 (221.0/21.0)
|   |   |   Hue > 5
|   |   |   |   Lightness <= 5: 1 (15.0)
|   |   |   |   Lightness > 5: 2 (5.0)
|   Saturation > 3
|   |   Hue <= 3: 1 (34.0)
|   |   Hue > 3
|   |   |   CIE76 <= 6.04: 2 (5.0)
|   |   |   CIE76 > 6.04: 1 (5.0)
CIE76 > 8.7
|   Lightness <= 17
|   |   Saturation <= -7
|   |   |   Lightness <= 10: 1 (18.0)
|   |   |   Lightness > 10
|   |   |   |   CIE76 <= 10.14: 2 (5.0)
|   |   |   |   CIE76 > 10.14: 1 (2.0)
|   |   Saturation > -7: 1 (105.0)
|   Lightness > 17: 2 (10.0)
```

Figure 18. Rule system for the image 550 and the expert 1.

The success rates of the rule-based system are always higher than 85%, with image 449, where this exceeds 91.5%, being especially notable. The precision and recall rates, except in image 448 for expert 1, are always higher than 0.83. The inconsistency rates we saw for all the experts would make it impossible to obtain a rules system directly from the analysis of their behaviour with performance of 100%, as the rule-based system, to a certain extent, models this inconsistency by inferring the rules directly from the data obtained for the experts themselves.

We can assume that the result of the machine-learning test reinforces the conclusion reached in the previous section on the existence of regular patterns in the quality judgments of the experts, that these patterns are based on visual analysis of perceptual colour properties, and that it is possible to generate a model that represents these regular patterns through the combined use of easily computable metrics and colour perceptual attributes, such as CIE76, CIE00 or HSL. Therefore, we understand that it is possible to generate a numerical model that, with a small set of variables, yields a relatively high success rate when compared to the error rates we found in the evaluations of the human experts participating in the experiment. The mathematical representation of this model would comprise a visual algorithm. In order to assess the complexity of a rules-based visual algorithm of these characteristics, we have analyzed the complexity of the decision trees. Except in the case of the first image, the resulting sizes are small, as the number of rules oscillates between 13 and 19. In the first image, they oscillate between 33 and 40. Therefore, the visual algorithm would really be efficient with the computing power currently available to us.

Next we performed three tests of the consistency of the machine-learning algorithm applied, using the data for expert 4, the most consistent, and the image in which the algorithm provided the highest success rate, 448.

a) Test of swapping the CIE76 metric with MCIE76.

We confirmed the variation in the performance of the algorithms if we use the samples taken of the images themselves (MCIE76 metric) instead of the colour patches from the control cards. As we saw above, their correlations are not as high as between metrics CIE76 and CIE00. The results are very similar to those obtained using the patches from the card. This conclusion is important, as we have to consider that in a quality evaluation system, it is not at all efficient to use samples of the documents instead of standardized colour cards, given the excessive amount of work time involved in taking samples and obtaining the standardized colour values.

Total images	Correctly classified		Incorrectly classified		Precision		Recall		Number of rules
728	Total	%	Total	%	Valid	Invalid	Valid	Invalid	35
	671	92.17	57	7.83	0.901	0.945	0.948	0.895	

Table XVI. Rule system performance using metric MCIE76.

b) Application of algorithm C4.5 exclusively for metrics CIE76 and CIE00 in isolation.

The results obtained by inferring the rules only from the CIE metrics in isolation are not acceptable, as the success rate is very low, 66.4% for CIE76 and 61.5% for CIE00, which was to be expected after observing the high degree of overlap between the data for the valid and invalid images according to these metrics.

c) Validation of the rule-based system obtained with a set of images not used to infer the rules, and the performance of the inducted rule-based system on data for images evaluated with different criteria.

The method of validating the resulting rules we used in the previous experiments can even involve in the validation process the use of some records already employed to generate the rules. This is because of the compensation of images that we included to bring the classes of valid and invalid images closer together. To avoid the problem of bias towards optimum results that this practice might represent, we proceeded to redo the experiment, applying algorithm C4.5 with a compensation system that does not involve including duplicate data records to balance the percentages of the two classes. Due to the disparity between the number of images considered valid and invalid for all the experts and images, we proceeded to bring the two classes closer together by creating a consolidated file with the records for the three images for expert 4, without including compensation. This file contains a total of 903 records, 223 of them corresponding to valid images and the remaining 680 to invalid ones. We are aware that by doing this we are mixing different criteria, since, as we saw earlier, the experts do not apply exactly the same criteria to the three images due to their differences in iconic

content; therefore the performance of the algorithm should be lower, and we should assume a reduction in the performance of the rule-based system. But with this test, we can see, at the same time, how the use of different criteria by the human evaluators we want to emulate influences the performance of the visual algorithm obtained automatically. Based on the consolidated file, we then created three files with the following distributions of records:

- File A. With the records of the 223 valid images and 227 invalid images.
- File B. With the records of the 223 valid images and 226 invalid images not included in the other two files.
- File C. With the records of the 223 valid images and the remaining 226 invalid images not included in the other two files.

To validate, we used the crossed validation method, which, by not acting on records with duplicates, in no case employs records used for inferring the rules in the validation process.

The resulting success rates are the following:

	Correctly classified	Valid images recall	Invalid images recall	Valid images precision	Invalid images precision
A	85.56	0.852	0.859	0.856	0.855
B	74.4	0.735	0.752	0.745	0.742
C	75.5	0.785	0.726	0.738	0.774

Table XVII. Rule system performance for files A, B and C.

The results become less successful, with an important factor being the lack of uniformity in the criteria applied to each of the images. Nevertheless, even with this limitation, the success rate is higher than 74% in all cases.

We must reflect on the disparity of results between the rules obtained for each image and by each expert. This disparity suggests that the criteria applied to value judgments vary according to the image motif and the expert. Both types of inconsistency are a problem for quality evaluation systems based on human experts. For this reason, studies are needed that address in greater depth how the type of image motif influences the perception of quality and the factors that cause lack of consistency between evaluators. The methods of analysis we used for this study can be used to detect and analyze this type of problem.

4. Conclusions

Quality control systems for heritage digitization must consider the performance of the quality measurement parameters, not only at the physical level but also at the overall perception level, modelling to the extent possible the complex interactions that take place between the image quality attributes at this level. A perceptual model involves knowledge that must be

obtained through experimentation with human quality experts with sufficient training in the objectives of the projects. These experiments run into the problem of inter- and intra-evaluator inconsistency, which must be measured beforehand.

We conclude that it is not possible to talk about continuous acceptance ranges for the metrics habitually considered in quality systems in colour and in the use of these metrics on an isolated basis, and therefore that it is necessary to investigate more complex models. In this study, we attempted to obtain a model based on a rule-based system with high performance for the case considered in the experiment presented employing the CIE76 and CIE00 metrics along with the HSL colour perceptual attributes. The detection of regular patterns of values for these attributes in the zone of overlap between images considered valid and invalid by the experts leads us to consider that this combination of attributes and metrics might be suitable for objectively measuring the subjective appreciation of perceptual proximity with a relatively high degree of success, which will always be limited by the errors committed by the human expert evaluators in their evaluation work.

We used machine-learning algorithm C4.5 in an attempt to obtain a rule-based system that would enable modelling of these behaviour patterns and which, therefore, could be applied to emulate the human experts with a high degree of efficacy. The results indicate that it is possible to emulate the scoring process of the expert with efficacy rates above 85% by these means. The percentage of errors committed by the experts was estimated at between 10.87% and 20%, and therefore we can consider their success rates comparable to those of the created system. Given the variability of inter- and intra-evaluator criteria detected, it is not possible to generalize a single model for the entire set of evaluators, although it can be assumed that after a long enough period for training and agreeing on the results, it would be possible to improve this inconsistency enough to generate a single, highly efficacious model.

References

Bureau Metamorfoze. Koninklijke Bibliotheek (National Library of the Netherlands) (2007). Metamorfoze Preservation Imaging Guidelines draft. Retrieved from http://www.metamorfoze.nl/sites/metamorfoze/files/bestanden/richtlijnen/guidelinespijune07.pdf Accessed 12 June 2011.

Charrier, C., Lézoray, O., & Lebrun, G. (2012). Machine learning to design full-reference image quality assessment algorithm. *Signal Processing: Image Communication*, 27(3), 209–219.

Dormolen, H. (2010). *Metamorfoze Preservation Imaging Guidelines*. Test Version 0.8. July 2010. Retrieved from http://www.metamorfoze.nl/sites/metamorfoze/files/bestanden/richtlijnen/Metamorfoze%20Preservation%20Imaging%20Guidelines_Version_0.8_July_2010.pdf Accessed 12 June 2011 Accessed 12 June 2013.

Engeldrum, P. G. (1995). A framework for image quality models. *Journal of Imaging Science and Technology*, 39(4), 312-318.

Engeldrum, P. G. (2004). A Theory of Image Quality: The Image Quality Circle. *Journal of Imaging Science and Technology*, 48(5), 446–456.

FADGI- Still Image Working Group (2010). *Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files. For the Following Originals - Manuscripts, Books, Graphic Illustrations, Artwork, Maps, Plans, Photographs, Aerial Photographs, and Objects and Artifacts*. Retrieved from http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf Accessed 12 June 2013.

Fairchild, M. D. (2004). *Color Appearance Models: CIECAM02 and Beyond*. In IS&T/SID 12th Color Imaging Conference. Tutorial T1A, 11/9/04. Retrieved from http://www.cis.rit.edu/fairchild/PDFs/AppearanceLec.pdf Accessed 12 December 2013.

Frey, F., & Reilly, J. (1999). *Digital Imaging for Photographic Collections: Foundations for Technical Standards*. Rochester, NY: Image Permanence Institute.

Frey, F., & Reilly, J. (2006). *Digital Imaging for photographic collections: foundations for technical standards*. (2nd ed.) Rochester, NY: Image Permanence Institute.

ISO 20462-1:2005 (2005a). Photography Psychophysical experimental methods for estimating image quality —Part 1: Overview of psychophysical elements.

ISO 20462-2:2005 (2005b). Photography -- Psychophysical experimental methods for estimating image quality — Part 2: Triplet comparison method.

ISO 11664-4:2008 (CIE S 014-4/E:2007) (2007). Colorimetry -- Part 4: CIE 1976 L*a*b* Colour space.

ISO 12646:2008 (2008). Graphic technology -- Displays for colour proofing -- Characteristics and viewing conditions.

ISO 3664:2009 (2009). Graphic technology and photography - Viewing conditions.

ISO 20462-3:2012 (2012). Photography -- Psychophysical experimental methods for estimating image quality — Part 3: Quality ruler method.

Lee, H. (2005). *Introduction to Color Imaging Science*. Cambridge: Cambridge University Press.

Luo, M. R., Cui, G., & Rigg, B. (2001). The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research & Application*, 26(5), 340–350.

Martens, J. B. (2002). Multidimensional modeling of image quality. *Proceedings of the IEEE*, 90 (1), 133–153.

Martínez, C., & Muñoz, J. (2002). Digitalización del patrimonio fotográfico e investigación: la metodología empleada para la reproducción digital de la colección de placas de vidrio de colodión húmedo, custodiada en el Museo Nacional de Ciencias Naturales –Consejo Superior de Investigaciones Científicas- (MNCN-CSIC). In *Actas de las Primeras Jornadas sobre Imagen, Cultura y Tecnología, Madrid, Spain, 2002,* Madrid: Universidad Carlos III de Madrid, 99-120.

Nationaal Archief (2010). *Digitisation of photographic materials. Guidelines*. September 2010. Retrieved from http://www.nationaalarchief.nl/sites/default/files/docs/guidelines_digitisation_photographic_materials.pdf Accessed 12 December 2013.

Narwaria, M., Lin, W. & Cetin, A. E. (2012). Scalable image quality assessment with 2D mel-cepstrum and machine learning approach. *Pattern Recognition*, 45(1), 299-313.

Pellacini, F., Ferwerda, J.A. & Greenberg, D.P. (2000). Toward a psychophysically-based light reflection model for image synthesis. *Proceedings of the ACM - SIGGRAPH 2000*, 55–64.

Puglia, S., Reed, J., & Rhodes, E. (2004). *U.S. National Archives and Records Administration (NARA) Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images*. Retrieved from http://www.archives.gov/preservation/technical/guidelines.pdf.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Robledano, J. (2011a). Mejora del rango dinámico en la digitalización de documentos desde una perspectiva patrimonial: evaluación de métodos de alto rango dinámico (HDR) basados en exposiciones múltiples. *Revista española de Documentación Científica*, 34 (3), 357-384.

Robledano, J. (2011b). Twenty-five years of digital conversion. Current situation. In Internacional Conference. *Thirty Years of Photographic Conservation Science,* June, 2011, Logroño (La Rioja), Spain. Retrieved from http://e-archivo.uc3m.es/handle/10016/16579 Accessed 12 December 2013.

Ruiz, P. (2006). Sistemas de control de calidad para la digitalización. In *Actas de las IX Jornadas Antoni Varés, Imatge i Recerca*. Girona: CRDI, 61-84.

Still Image Working Group (2010). *GAP Analysis. Updated 01/12/2010*. Retrieved from http://www.digitizationguidelines.gov/stillimages/documents/Gap_Analysis.pdf Accessed 12 December 2013.

Tchan, J., Thompson, R.C., Manning, A. (1999). A computational model of print-quality perception. *Expert Systems with Applications*, 17(4), 243–256.

Williams, D. (2002). Image quality metrics. *RLG Diginews*, 4(4). Retrieved from http://www.worldcat.org/arcviewer/1/OCC/2007/08/08/0000070511/viewer/file1806.html Accessed 17 January 2014.

Williams, D. (2003). Debunking of Specsmanship: Progress on ISO/TC42 Standards for Digital Capture Imaging Performance. In *IS&T's 2003 PICS Conference*, 77-81. Retrieved from http://www.i3a.org/wp-content/uploads/2008/03/debunking_specsmanship.pdf Accessed 17 January 2014.

Williams, D. (2010). *Imaging Science for Archivists*. Retrieved from http://www.docstoc.com/docs/50793406/Imaging-Science-for-Archivists---101-Don-Williams---Image Accessed 17 January 2014.

Witten, I. H., & Frank, E. (2005). *Data Mining. Practical Machine Learning Tools and Techniques* (2nd Ed.) San Mateo, CA: Morgan Kaufmann Publishers.

Wueller, D., Dormolen, H., & Jansen, V. (2009). *Universal Test Target Technical Specification*. Retrieved from http://www.universaltesttarget.com/download/UTT%20technical%20specs%20v1.1.pdf Accessed 17 January 2014.

Zhiqing, W., & Yang, T. (1999). Building a rule-based machine vision system for defect inspection on apple sorting and packing lines. *Expert Systems with Applications*, 16(3), 307–313.

Zhou, W., Bovik, A.C., & Ligang L. (2002). Why is image quality assessment so difficult?. *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, Volume 4, 13-17 May 2002, IV-3313 - IV-3316 doi: 10.1109/ICASSP.2002.5745362.