Revista Española de Documentación Científica
39(2), abril-junio 2016, e128
ISSN-L:0210-0614. doi: http://dx.doi.org/10.3989/redc.2016.2.1249

ESTUDIOS / *RESEARCH STUDIES*

Aproximación experimental al uso de métricas objetivas para la estimación de calidad cromática en la digitalización de patrimonio documental gráfico

Jesús Robledano-Arillo*, Valentín Moreno-Pelayo**, José Manuel Pereira-Uzal***

*Departamento de Biblioteconomía y Documentación. Universidad Carlos III de Madrid.
**Departamento de Informática. Universidad Carlos III de Madrid.
***DigitalHeritage (www.jpereira.net)
Correo-e: jroble@bib.uc3m.es, vmpelayo@inf.uc3m.es, info@jpereira.net

Recibido: 30-12-2014; 2ª versión: 13-07-2015; Aceptado: 15-07-2015.

Cómo citar este artículo/Citation: Robledano-Arillo, J.; Moreno-Pelayo, V.; Pereira-Uzal, J. M. (2016). Aproximación experimental al uso de métricas objetivas para la estimación de calidad cromática en la digitalización de patrimonio documental gráfico. *Revista Española de Documentación Científica*, 39(2): e128. doi: http://dx.doi.org/10.3989/redc.2016.2.1249

Resumen: Se abordan de una forma crítica diferentes aproximaciones aplicables para la realización de modelos de sistemas de control de calidad automatizado de imágenes digitales en proyectos de digitalización de fondos fotográficos con valor histórico-cultural. Tras la realización de un experimento psicométrico con cuatro expertos humanos se concluye que no es posible utilizar con un buen rendimiento los modelos simplistas de uso común basados en rangos de aceptación continuos sobre mediciones de color tomadas de forma aislada. Nuestra investigación demuestra que un modelo basado en un sistema de reglas obtenidas por aprendizaje automático que emplee las métricas CIE 1976 o CIEDE 2000, junto con los atributos perceptuales de color matiz, saturación y luminosidad, emula a los expertos humanos en calidad de imagen con un alto grado de eficacia, por encima del 85%.

Palabras clave: Digitalización de documentos; fotografía; evaluación de calidad; aprendizaje automático; algoritmos visuales.

Experimental approach to the use of objective metrics for estimating chromatic quality in the digitization of graphical documents

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. After conducting a psychometric experiment with four human experts, we demonstrate that it is not possible to talk about commonly used, simplistic models based on continuous acceptance ranges for colour metrics on an isolated basis. This study demonstrates that a model based on a rule-based, machine-learning system employing metrics (CIE 1976 or CIEDE 2000) along with the colour perceptual attributes of hue, saturation and lightness, emulates the image quality experts with a high degree of efficacy, above 85%.

Keywords: Document digitization; photography; image quality assessment; machine learning; visual algorithms.

Copyright: © 2016 CSIC. Este es un artículo de acceso abierto distribuido bajo los términos de la licencia Creative Commons Attribution-Non Commercial (by-nc) Spain 3.0.

1. INTRODUCCIÓN

En el contexto de las digitalizaciones patrimoniales de fotografías y otros documentos con valor gráfico se ha venido imponiendo una perspectiva estricta de calidad que concibe las imágenes digitales como representaciones fieles a nivel físico y perceptual: las imágenes deben representar con fidelidad las características físicas de los documentos físicos originales y su apariencia ante unas condiciones de percepción determinadas durante el proceso de captura digital. Sólo así podrán ser usadas para las funciones de custodia, conservación, reproducción, análisis, estudio y divulgación a las que deben dar soporte, dentro de unos criterios éticos que no amparan el cambio de las características plásticas ni la reinterpretación de los mensajes icónicos y plásticos (Martínez y Muñoz, 2002; Ruiz, 2006; Robledano, 2011a, 2011b). Esta perspectiva estricta tiene importantes implicaciones a la hora de plantear un procedimiento de control de calidad de las digitalizaciones, pues su aplicación introduce la necesidad de manejar dos planos: un plano físico y un plano perceptual.

De acuerdo al primero, el nivel de calidad puede ser medido objetivamente de forma sencilla aplicando atributos físicos de la imagen que han sido ampliamente estudiados en las últimas décadas desde los campos de la ingeniería de la imagen y la ciencia y tecnología del color, tales como la capacidad resolutiva obtenida en el registro de información gráfica, el error en la codificación del color, el rango dinámico, la OECF (Opto-Electronic Conversion Function), etc.; así como midiendo el grado de afectación de una serie de distorsiones de la señal digital que pueden afectar al rendimiento de los atributos, como ruido, aberraciones cromáticas, distorsiones geométricas, artefactos de compresión, etc. Ha habido incluso varios intentos de sistematizar estas características desde el contexto de la digitalización patrimonial de fondos y colecciones culturales de diversa tipología (Frey y Reilly, 1999 y 2006; FADGI, 2010). A partir de un conjunto de atributos físicos preseleccionados puede construirse un modelo de calidad multidimensional que permite computar la calidad de la imagen digital de un documento original que ha sido digitalizado junto a una o varias cartas de referencia con respecto a su original físico correspondiente. Se han venido usando diferentes modelos multidimensionales para computar la calidad a partir de las medidas obtenidas sobre los atributos, como el Generalized Weighted Mean hypothesis o las métricas de Minkowski (Engeldrum, 1995). La calidad en este tipo de modelos se puede aproximar como una función que calcula la distancia euclídea de las imágenes degradadas con respecto a una imagen ideal en un espacio n dimensional, siendo las dimensiones los atributos incluidos en los experimentos (ecuación I).

Ecuación I. Calidad (C) entendida como distancia euclidiana entre una imagen digital (x) y su imagen referente ideal (y), a partir de sus atributos (i), ponderados mediante sus coeficientes de ponderación (p).

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

De acuerdo al plano perceptual, la calidad de la imagen es la posibilidad de generar visualizaciones o reproducciones a partir de ésta que provoquen al usuario una percepción global similar a la que tendría si observara el documento original ante unas condiciones de observación determinadas y controladas, y sin distorsiones de ninguna clase. La apreciación global de calidad a nivel perceptual es un proceso subjetivo que se realiza comúnmente mediante la visualización por parte de un observador humano del documento físico junto a una reproducción o visualización de su correspondiente imagen digital. El observador humano tratará de cuantificar el grado en que la imagen digital se aleja perceptualmente de su correspondiente original, en unas condiciones de visualización normalizadas de acuerdo a estándares (ISO, 2008, 2009). La introducción de un evaluador humano es muy costosa, por lo que se hace preciso para muchos proyectos de digitalización masivos crear sistemas de control de calidad automáticos que sustituyan al observador experto humano en la fase de cotejo de calidad, pero que no mermen el alto rendimiento que un observador humano experimentado tiene a la hora de evaluar la proximidad perceptual entre el original y su correspondiente imagen digital.

Dada la facilidad de computación de los atributos y distorsiones de tipo físico, una línea importante de investigación sobre la forma de desarrollar estos sistemas ha sido el intento de conectar los niveles de rendimiento físico y perceptual, de manera que se pueda derivar automáticamente la calidad global de una imagen a nivel perceptivo mediante el uso de medidas de tipo físico fácilmente computables, trabajando con un reducido número de atributos y rangos de valores, y con procesos altamente eficientes. Se emplea usualmente el término algoritmo visual para denominar este tipo de modelos matemáticos. El problema reside en que no es sencillo poder derivar directamente la fidelidad perceptual a partir de la fidelidad física. Muchos esfuerzos a la hora de crear un algoritmo visual robusto a nivel de la percepción subjetiva humana de la calidad global a partir de atributos de tipo físico han fracasado por no haber considerado suficientemente la multiplicidad de elementos y complejas interrelaciones que subyacen en este fenómeno en un modelo de calidad suficientemente exhaustivo (Engeldrum, 2004;

Zhou y otros, 2002). El rendimiento de este tipo de aproximaciones se resiente por diversos motivos, como la falta de linealidad de la percepción humana de los problemas de calidad, el uso de atributos que carecen de un grado fuerte de correspondencia con la apreciación perceptual de calidad (Engeldrum, 2004), por tratar los atributos de forma independiente sin considerar que son mutuamente interactivos (Lee, 2005), o por no haber incorporado la influencia de una serie de factores subjetivos que condicionan la interpretación visual de la imagen y que han sido ampliamente estudiados (Fairchild, 2004).

La aplicación de métodos de escalado multidimensionales que permiten analizar las complejas interacciones subyacentes en los atributos de calidad de las imágenes ha sido explicado por Lee (2005), quien refiere algunos de los que han obtenido descriptores físicos para atributos psicofísicos (Martens, 2002; Pellacini y otros, 2000). También se ha descrito la aplicación de métodos de aprendizaje automático a través de los cuales se pueda llegar a inferir los atributos de calidad determinantes y sus modelos de interrelación para la automatización de sistemas de control de calidad de imágenes, pero en ámbitos alejados del contexto de actividad donde nos vamos a centrar y empleando bases de datos gráficas de experimentación, tales como LIVE o TID2008, cuyas características se alejan de las del tipo de objeto patrimonial al que dirigimos nuestra investigación. En esta línea se han propuesto varios métodos dentro de lo que se suele denominar como Machine Learning-based Image Quality Measure (MLIQM), tal como el descrito por Charrier y otros (2012), basado en aplicar un sistema de aprendizaje automático para clasificar imágenes y que trata de superar las limitaciones de métodos cercanos por su planteamiento. Esta aportación es de alcance limitado para los objetivos patrimoniales que proponemos.

En el contexto de los objetos gráficos patrimoniales los sistemas automatizados de calidad de las digitalizaciones se han venido basando fundamentalmente en pruebas encuadrables dentro del nivel físico, usándose exclusivamente un conjunto limitado de atributos de esta tipología, para los que se establecen unos rangos de aceptación de valores previamente determinados. Si atendemos a los principales trabajos que se han publicado sobre esta cuestión dentro del campo del patrimonio documental, podemos concluir que la mayoría de ellos se ha encaminado a la identificación y propuestas de métricas de medida de atributos exclusivamente físicos, pero sin profundizar en un modelo perceptual de calidad global de la imagen que guie a la hora de establecer los rangos de aceptación sistemáticos en el rendimiento de estos atributos y sus complejas interrelaciones durante el acto de percepción (Williams, 2002, 2003 y 2010; Puglia y otros, 2004; FADGI, 2010; Still Image Working Group, 2010; Dormolen, 2012; Nationaal Archief, 2010).

Nuestro trabajo se centra en el intento de establecer una vía de trabajo válida para la creación automática de algoritmos visuales altamente eficientes que puedan ser usados en sistemas de control de calidad de imágenes digitales provenientes de la digitalización de obras de tipo gráfico, y que permitan superar las limitaciones de los sistemas basados en modelos multidimensionales que usan un conjunto predefinido de atributos de calidad junto a sus rangos de valores de aceptación. Debido a la amplitud de este objetivo, abordaremos exclusivamente el uso de atributos de color. Tratamos de demostrar que es posible modelar los juicios perceptuales de valor de un experto o de un conjunto de expertos evaluadores humanos, en lo que respecta a la proximidad perceptual en color entre una imagen digital y su correspondiente original físico, mediante un algoritmo visual computable de forma eficiente que se base en el uso conjunto de métricas de medida de color estandarizadas y de atributos perceptuales del color. Ante la complejidad de las interacciones entre los atributos de color que se producen en el acto perceptivo se hace precisa la automatización del proceso de obtención del algoritmo visual. Para ello proponemos la aplicación de un método de aprendizaje automático basado en la inducción de reglas que no requiera predefinir de antemano los atributos de calidad más determinantes y sus rangos de aceptación y que pueda trabajar sobre un conjunto de datos obtenidos de procesos reales de evaluación humanos que se quieran modelar. En este trabajo hemos aplicado el algoritmo de aprendizaje automático a los datos obtenidos de un conjunto experimental de imágenes previamente evaluadas por un grupo de expertos humanos en evaluación de imágenes.

2. METODOLOGÍA

2.1. Fase I. Prueba de evaluación con expertos humanos

La prueba ha consistido en la emulación de un proceso real de evaluación de calidad con expertos humanos, aplicando unas condiciones de contexto de evaluación ideales, de acuerdo a la normativa estandarizada para realizar procesos de evaluación de calidad mediante el cotejo de los originales físicos con las imágenes digitales correspondientes en pantalla: ISO 3664 (ISO, 2009), 12646 (ISO, 2008), 20462-1:2005 (ISO, 2005a), 20462-2:2005 (ISO, 2005b) y 20462-3:2012 (ISO, 2012). Se han usado tres imágenes fotográficas sobre papel representativas del tipo de documentos que está presente en muchos fondos fotográficos: materiales positivos fotográficos en color modernos en acabado brillo y mate, y materiales fotográficos antiguos con iluminación a mano mediante tintas. Hemos elegido diversos motivos icónicos, sobre la premisa de que el motivo de la imagen influye en la percepción de cali-

dad. Hemos elegido figuras humanas (imagen nº 448) y paisajes con iconos usuales -cielo, nubes, hierba, bosque, agua- (imágenes nº 449 y 550).

Hemos creado a continuación los másteres, digitalizando directamente las imágenes originales junto a una carta de color colorchecker con una cámara fotográfica digital réflex y aplicando gestión a través de perfiles de color ICC personalizados, para conseguir así imágenes con alta fidelidad de color y contraste a nivel colorimétrico y densitométrico. A partir de los másteres se creó una serie de entre 303 y 300 imágenes degradadas por cada original físico, mediante la edición de sus valores perceptuales HSL: Matiz/Hue (H), Saturación (S) y Luminosidad (L). Así se ha creado una secuencia de degradación que contempla una escala suficientemente amplia de cambios perceptibles en estas tres variables descriptivas de color. Para ello las imágenes han sido transformadas al sistema de color HSL y degradadas progresivamente en estas tres variables, respectivamente, desde -20 a 19 para Matiz (en una escala que va desde -100 a +100), desde -39 a +39 para Saturación (en una escala que va desde -100 a +100) y desde -20 a 20 valores para Luminosidad en una escala que va desde -100 a 99. Se han generado asimismo imágenes repetidas, con la finalidad de poder medir el grado de consistencia en las evaluaciones, analizando cómo varía su criterio selectivo a lo largo del tiempo de la prueba, en su caso, y poder determinar la probabilidad de respuesta aleatoria de los evaluadores durante toda la prueba. Las imágenes repetidas se han repartido a lo largo de cada serie de imágenes a evaluar.

Se han registrado automáticamente los datos de las imágenes a evaluar aplicando diferentes métricas de cotejo de diferencia de color y de imagen entre las fotografías originales y los másteres digitales y sus degradaciones, de las que hemos seleccionado sólo dos de ellas para las pruebas que presentamos en este trabajo: CIEDE 2000 -CIE00- (Luo y otros, 2001) y CIE 1976 L*a*b* colour-difference formula -CIE76- (ISO, 2007). Para los cálculos de las diferencias de color entre las imágenes físicas digitalizadas y las imágenes digitales de las series de degradados hemos usado todos los parches de la carta colorchecker.

La evaluación visual de los expertos se realizó sobre la percepción de las imágenes digitales reproducidas en el monitor, por lo que se hizo preciso controlar minuciosamente todos los elementos que conforman el flujo de visualización, que son, además de la imagen digital: la calibración y perfilado ICC del monitor; la conversión desde el espacio de color de la imagen al espacio de color del monitor hecho por el Gestor del Color (CMS) del sistema operativo; la calidad del monitor y de sus condiciones de entorno de visualización; la calidad de la cabina de visualización de los originales físicos y de sus condiciones de visualización. El interfaz de pantalla fue diseñado con el programa Adobe Bridge de forma que en ésta sólo aparecía la imagen en proceso de evaluación y en su margen izquierdo una fina tira con las imágenes del grupo que servían a los evaluadores para ir seleccionando la siguiente imagen a visualizar y poderse mover por el lote. En la cabina de visualización se ubicó la carta colorchecker usada para hacer los másteres y el propio original con una ubicación muy similar a la que presentaban las imágenes de la prueba. La intensidad del color gris de fondo de la pantalla se hizo coincidir con el de la cabina. Los expertos pudieron asignar a cada imagen, según la calidad detectada, una puntuación basada en una escala de 3 valores: 1 (la imagen no pasaría un control de calidad profesional que mide la proximidad en apariencia de color y contraste entre una imagen en pantalla y una imagen en papel); 2 (la imagen pasaría el control de calidad pero con un criterio no muy riguroso); y 3 (la imagen pasaría el control de calidad con un criterio riguroso). Con ánimo de simplificar la primera aproximación analítica que hacemos en este estudio, hemos unido los valores 2 y 3, de manera que trabajaremos sólo con dos clases de calidad: imagen válida e imagen inválida.

Se seleccionaron cuatro expertos que cumplían la condición de ser profesionales con una dilatada experiencia en los sectores de la fotografía profesional y Artes Gráficas y en trabajos de evaluación de calidad de imágenes digitales (8, 14, 15 y 16 años de experiencia laboral de evaluación). El equipo de expertos fue instruido con el tiempo suficiente como para entender el tipo de evaluación de calidad que se requiere en el campo del patrimonio documental.

2.2. Fase II. Análisis de datos

Con los datos recogidos se realizaron dos tipos de análisis:

1) Análisis de coherencia en los juicios de calidad de cada evaluador.

Sus objetivos han sido dos: detectar y estimar porcentualmente los errores por falta de consistencia en las evaluaciones de los expertos humanos participantes en la prueba, y poder comparar los porcentajes de error de los expertos con el del sistema de reglas que obtengamos posteriormente mediante aprendizaje automático. Hemos aplicado dos parámetros que nos permiten medir el grado de consistencia en las evaluaciones de cada evaluador humano (intra-evaluadores) y entre los evaluadores (inter-evaluadores).

a) Error de consistencia intra-evaluadores.

Este tipo de falta de consistencia es indicativa de la aplicación de procesos de evaluación aleatorios en algunos momentos de la prueba o de cambios en los criterios de calidad que se emplean a lo largo de ésta. La consistencia ha sido medida a través de las imágenes repetidas insertadas en las series. Para su cálculo se han sumado, para cada exper-

to, todos los errores de consistencia ocurridos a lo largo de las tres series de imágenes, y el número total de imágenes repetidas, y se ha hallado el porcentaje que representan los primeros con respecto del total de aquellas. Se ha entendido un error de consistencia como una diferencia en la asignación de valoración a las imágenes repetidas idénticas.

b) Consistencia inter-evaluadores.

La falta de consistencia en este nivel se debe principalmente al uso de diferentes criterios o grados de exigencia durante la evaluación. Hemos aplicado tres indicadores:

- Grado de exigencia de los expertos. Para su cálculo se ha hallado el porcentaje de imágenes seleccionadas como válidas de entre el total de imágenes evaluadas.
- Grado de consistencia entre expertos con respecto a la coincidencia de valoraciones en las mismas imágenes. Para su cálculo hemos medido el porcentaje de coincidencia entre cada par de expertos en los tres grados de valoración permitidos en la prueba. Se trata de medir en qué proporción cuando un experto ha asignado un determinado valor a las imágenes de la serie el resto de expertos han coincidido con él.
- Grado de coincidencia en las valoraciones de todas las imágenes por parte de los cuatro evaluadores. Representado por la suma de imágenes donde todos los expertos han coincido en el mismo valor y en la de las imágenes donde no ha habido esta coincidencia.

2) Análisis de regularidades en el comportamiento de las variables perceptuales de color HSL y de las métricas de diferencia de color CIE en los juicios de calidad de los evaluadores.

Hemos tratado de detectar si existen o no patrones regulares intra e inter-evaluadores en la dispersión de los valores de los diferentes atributos que expliquen el criterio de calidad que están aplicado los expertos, y cuáles son los atributos perceptuales del color que mejor permiten modelar el comportamiento de los expertos. La existencia de estos patrones facilitaría el trabajo de obtención de algoritmos visuales, a partir de los cuales se podrían generar sistemas de evaluación altamente eficientes y que se aproximen en precisión a los procesos de evaluación humanos. Tras analizar los resultados deberemos ser capaces de determinar si es viable generalizar modelos de calidad basados en unos rangos fijos de aceptación sobre las métricas de diferencia y atributos perceptuales de color considerados en este estudio.

Para ello hemos analizado los rangos de valores de aceptación (valoración de calidad 2 y 3) y rechazo (valoración de calidad 1) en las métricas y variables HSL para cada uno de los expertos e imágenes, intentando detectar alguna regularidad en ellos. Posteriormente hemos analizado, de forma comparativa, el comportamiento de los valores de las imágenes degradadas en las variables HSL con respecto a los valores de las métricas de diferencia de color en el grupo de las imágenes aceptadas como válidas y en el de las rechazadas.

2.3. Fase III. Aplicación de un método de aprendizaje automático para la obtención y validación de un algoritmo visual

La detección de patrones de comportamiento regulares en las variables analizadas ha reforzado la idea de la utilidad de aplicar técnicas de aprendizaje automático para la obtención de un algoritmo visual que modele los patrones de comportamiento de éstas con una alta capacidad predictiva. Hemos aplicado el algoritmo de inducción de reglas C4.5 (Quinlan, 1993) a través del software de aprendizaje automático Weka (Witten y Frank, 2005), usando como umbral de confianza para la poda 0,25 y como número mínimo de casos por hoja 2. La validación se ha realizado mediante validación cruzada con 10 divisiones estratificadas. Hemos usado como atributos de instancia para la inducción de reglas la métrica CIE76 y los tres atributos perceptuales del color HSL (Matiz, Saturación, Luminosidad). Hemos utilizado todos los datos de las tres imágenes y los de dos de los expertos: el experto 1, por ser el más inconsistente, y el experto 4, por ser el más consistente. Los valores 2 y 3 se han continuado asimilando a una única clase, de forma que podamos manejar un atributo de clase de tipo binario. Se ha compensado al 50% el número de instancias positivas (imágenes válidas) y negativas (imágenes inválidas) para evitar la polarización del modelo hacia la clase más numerosa. La compensación se ha hecho repitiendo los registros de datos de las imágenes evaluadas positivamente.

3. RESULTADOS Y DISCUSIÓN

3.1. Consistencia de la respuesta de los expertos

a) Error de consistencia intra-evaluadores (véase Tabla I).

b) Consistencia inter-evaluadores.

- Grado de exigencia de los expertos (véase Figura 1).
- Grado de consistencia entre expertos con respecto a la coincidencia de valoraciones en las mismas imágenes.

Presentamos el promedio de los porcentajes obtenidos en las tres imágenes en varias tablas (véase Tabla II, Tabla III, Tabla IV y Tabla V).

- Grado de coincidencia en las valoraciones de todas las imágenes por parte de los cuatro evaluadores (véase Figura 2, Figura 3 y Figura 4).

Tabla I. Error de consistencia intra-evaluadores de los cuatro expertos

Experto 1	Experto 2	Experto 3	Experto 4
20%	15,22%	15,22%	10,87%

Figura 1. Grado de exigencia de los expertos, representado por el porcentaje de imágenes consideradas válidas para cada uno de los cuatro expertos

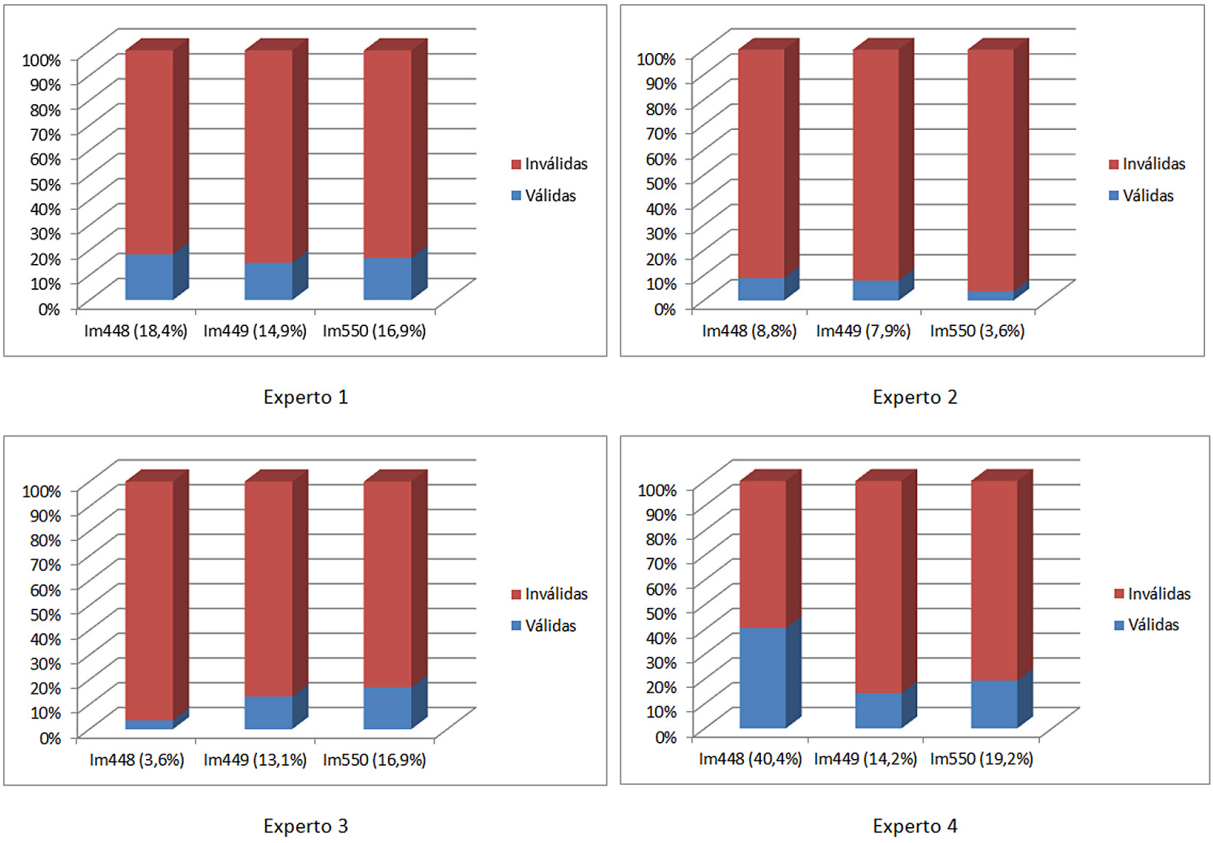


Tabla II. Promedio de % de coincidencia del experto 1 con respecto al resto de expertos

	Promedio de % de coincidencia en la asignación de 1	Promedio de % de coincidencia en la asignación de 2	Promedio de % de coincidencia en la asignación de 3
Experto 1 y Experto 2	95,97	14	0
Experto 1 y Experto 3	91,03	25,75	9,5
Experto 1 y Experto 4	82,87	38,6	25,57

Tabla III. Promedio de % de coincidencia del experto 2 con respecto al resto de expertos

	Promedio de % de coincidencia en la asignación de 1	Promedio de % de coincidencia en la asignación de 2	Promedio de % de coincidencia en la asignación de 3
Experto 2 y Experto 1	85,7	31,33	0
Experto 2 y Experto 3	88,97	21,8	0
Experto 2 y Experto 4	78,63	43,57	0

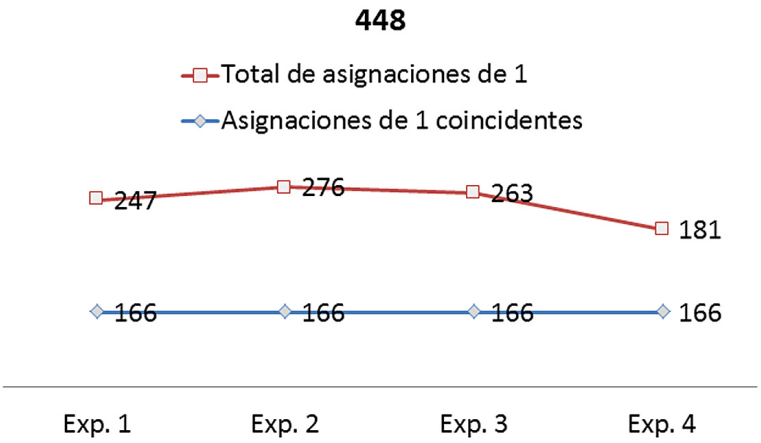
Tabla IV. Promedio de % de coincidencia del experto 3 con respecto al resto de expertos

	Promedio de % de coincidencia en la asignación de 1	Promedio de % de coincidencia en la asignación de 2	Promedio de % de coincidencia en la asignación de 3
Experto 3 y Experto 1	87,03	36,77	3,17
Experto 3 y Experto 2	95,5	13,97	0
Experto 3 y Experto 4	82,67	51,7	39,67

Tabla V. Promedio de % de coincidencia del experto 4 con respecto al resto de expertos

	Promedio de % de coincidencia en la asignación de 1	Promedio de % de coincidencia en la asignación de 2	Promedio de % de coincidencia en la asignación de 3
Experto 4 y Experto 1	91,73	34,9	14,3
Experto 4 y Experto 2	97,43	18,63	0
Experto 4 y Experto 3	95,5	32,26	13,26

Figura 2. Coincidencias de valoración en las imágenes de la serie 448 por todos los expertos



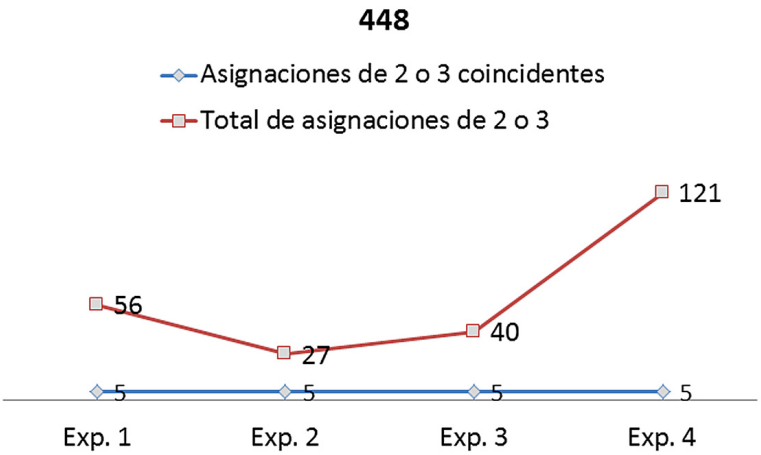
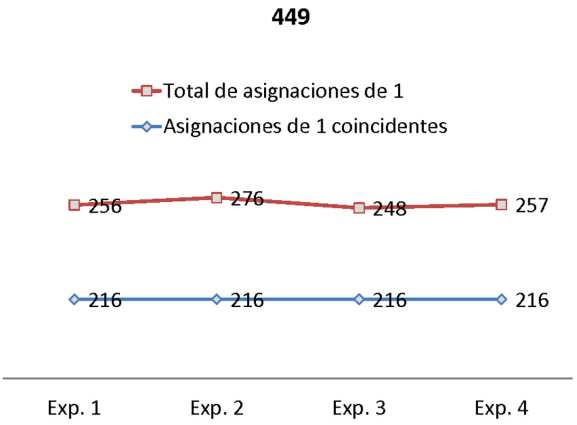


Figura 3. Coincidencias de valoración en las imágenes de la serie 449 por todos los expertos



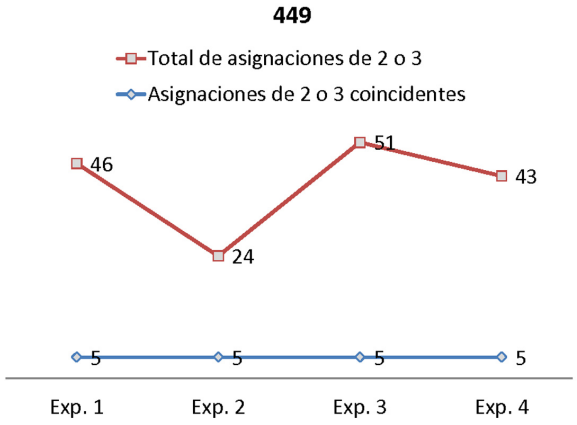
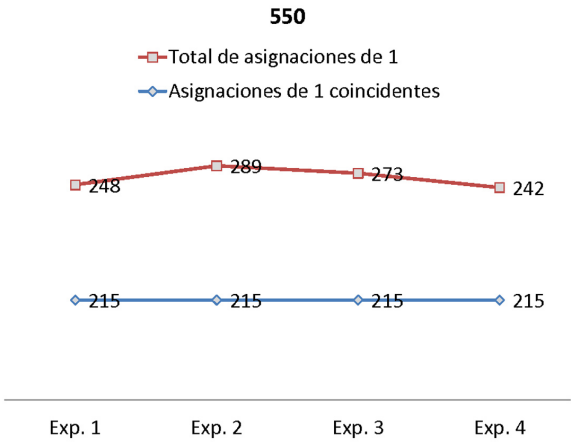
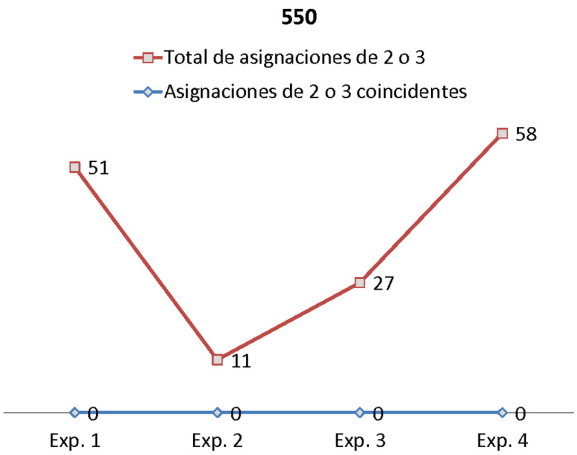


Figura 4. Coincidencias de valoración en las imágenes de la serie 550 por todos los expertos





Hay errores de consistencia en todos los expertos con una variabilidad de entre un 5 y un 10%. Los porcentajes de error no son muy altos, por lo que podemos descartar una respuesta aleatoria sostenida, asumiendo que los expertos han evaluado guiados por su percepción de calidad de las imágenes y aplicando un criterio de calidad. El experto más experimentado en evaluación de patrimonio documental, el número 4, es el más consistente por lo que parece que la formación de partida ha sido un factor que ha influido en el rendimiento en este parámetro. Parece que el grado de exigencia de los cuatro expertos es en general muy alto, pues los porcentajes, salvo en la imagen 448, están por debajo del 20%. Sólo parece haber disparidad de opiniones en la imagen 448, donde la diferencia entre porcentajes es alta, y para uno de los expertos (experto 2), el más exigente, en la imagen 550. Podemos decir que la coincidencia es en general baja por lo que no podemos hablar de uniformidad en el criterio de los cuatros expertos. Por ello, sin un período previo de puesta en común de criterio entre los expertos humanos participantes, un control de calidad dará siempre una tasa de consistencia baja, ofreciendo poca fiabilidad y coherencia.

Los grados de consistencia intra e inter-evaluadores indican la dificultad de conseguir porcentajes de eficacia muy altos en un algoritmo visual que modele su comportamiento de forma muy precisa, ya que el algoritmo modelará también las inconsistencias. Las inconsistencias a nivel intra-evaluadores son menores, por lo que es factible obtener unas tasas de rendimiento mayores obteniendo algoritmos individuales para cada experto. En un caso real habría que tratar de analizar el por qué se producen las inconsistencias, mejorando la formación de los expertos para poder aumentar

los niveles de consistencia antes de proceder a la obtención de un algoritmo visual, que pueda ser utilizado para el control automatizado de calidad, a través de este método.

3.2. Regularidades en el comportamiento de las variables perceptuales de color HSL y de las métricas de diferencia de color CIE en los juicios de calidad de los evaluadores

Presentamos en primer lugar una tabla con los valores límite en todas las métricas y variables perceptuales del color para cada imagen (Tabla VI), y a continuación esos datos pero referidos a cada experto (véase Tabla VII, Tabla VIII, Tabla IX y Tabla X). Estos datos ayudan a calibrar el alcance de los rangos de aceptación de las imágenes.

Presentamos los rangos de las variables utilizadas en el estudio en una serie de gráficos (Véase Figura 5, Figura 6, Figura 7 y Figura 8). Con ánimo de reducir la presentación y visualización de los datos hemos simplificado las métricas CIE basándonos en su correlación, eligiendo de las más correlacionadas sólo una de ellas, la métrica CIE76, que es una de las más empleadas en evaluación de calidad de color. También presentamos exclusivamente los datos de imagen 448, pudiendo considerar tras cotejar los datos de todas las imágenes, que se pueden generalizar estas conclusiones para las tres imágenes del estudio. Con la finalidad de evitar en los gráficos superposiciones de puntos hemos introducido perturbaciones aleatorias en los valores, consiguiendo de este modo una mejor visualización. Para ello hemos utilizado la función Jitter de la mencionada aplicación Weka, desplazando el cursor un 40%, aproximadamente, de su recorrido total.

En todas las métricas, salvando el extremo más inferior en el CIE76, el rango para las no válidas incluye el rango para las válidas, excepto algunas discontinuidades interiores. El solapamiento en los

Tabla VI. Valores máximo y mínimo de las métricas en las tres series de imágenes

	Valor	CIE76	CIE00	Matiz	Satur	Lum
448	Máx.	16,51	9,72	19	39	19
	Mín.	0,97	0,66	-20	-39	-20
449	Máx.	16,55	9,35	19	39	19
	Mín.	1,11	0,80	-20	-39	-20
550	Máx.	16,79	9,62	19	39	19
	Mín.	0,88	0,69	-20	-39	-20

Tabla VII. Rangos de valores para el experto 1

	Valor	CIE76	CIE00	Matiz	Satur	Lum
448	2 y 3	0,97 a 11,68	0,66 a 7,71	- 4 a 10	-31 a 8	-12 a 18
	1	1,04 a 16,51	0,68 a 9,72	-20 a 19	-39 a 39	-20 a 19
449	2 y 3	1,18 a 10,86	0,90 a 7,84	-5 a 9	-25 a 14	-3 a 19
	1	1,11 a 16,55	0,80 a 9,35	-20 a 19	-39 a 39	-20 a 18
550	2 y 3	0,88 a 10.74	0,69 a 7,84	-6 a 6	-23 a 8	-7 a 19
	1	0,88 a 16,79	0,69 a 9,62	-20 a 19	-39 a 39	-20 a 17

Tabla VIII. Rangos de valores para el experto 2

	Valor	CIE76	CIE00	Matiz	Satur	Lum
448	2 y 3	1,09 a 11,63	0,74 a 6,97	-2 a 11	-13 a 14	-7 a 13
	1	0,97 a 16,51	0,66 a 9,72	-20 a 19	-39 a 39	-20 a 19
449	2 y 3	1,11 a 7,58	0,80 a 5,91	0 a 4	-17 a 5	-7 a 13
	1	1,18 a 16,55	0,89 a 9,35	-20 a 19	-39 a 39	-20 a 19
550	2 y 3	0,88 a 8,47	0,69 a 6,70	0 a 6	-10 a 2	-3 a 13
	1	0,88 a 16,79	0,69 a 9,62	-20 a 19	-39 a 39	-20 a 19

Tabla IX. Rangos de valores para Experto 3

	Valor	CIE76	CIE00	Matiz	Satur	Lum
448	2 y 3	0,97 a 10,74	0,66 a 7,79	-2 a 8	-10 a 14	-7 a 19
	1	1,08 a 16,51	0,68 a 9,72	-20 a 19	-39 a 39	-20 a 16
449	2 y 3	1,18 a 10,86	0,90 a 7,84	-4 a 3	-13 a 9	-2 a 19
	1	1,18 a 16,55	0,89 a 9,35	-20 a 19	-39 a 39	-20 a 18
550	2 y 3	0,88 a 7,39	0,69 a 5,85	-1 a 2	-10 a 1	-11 a 13
	1	0,88 a 16,79	0,69 a 9,62	-20 a 19	-39 a 39	-20 a 19

Tabla X. Rangos de valores para el experto 4

	Valor	CIE76	CIE00	Matiz	Satur	Lum
448	2 y 3	0,97 a 13,88	0,66 a 8,32	-4 a 13	-35 a 14	-15 a 19
	1	1,08 a 16,51	0,68 a 9,72	-20 a 19	-39 a 39	-20 a 13
449	2 y 3	1,11 a 10,86	0,80 a 7,84	-4 a 6	-17 a 19	-7 a 19
	1	1,18 a 16,55	0,89 a 9,35	-20 a 19	-39 a 39	-20 a 18
550	2 y 3	0,88 a 10,221	0,69 a 7,42	-7 a 6	-17 a 2	-14 a 18
	1	0,88 a 16,79	0,69 a 9,62	-20 a 19	-39 a 39	-20 a 19

Figura 5. Distribución de imágenes inválidas (1) y válidas (2) en los rangos de valores de CIE76 para los cuatro expertos (en orden del 1 al 4). Imagen 448

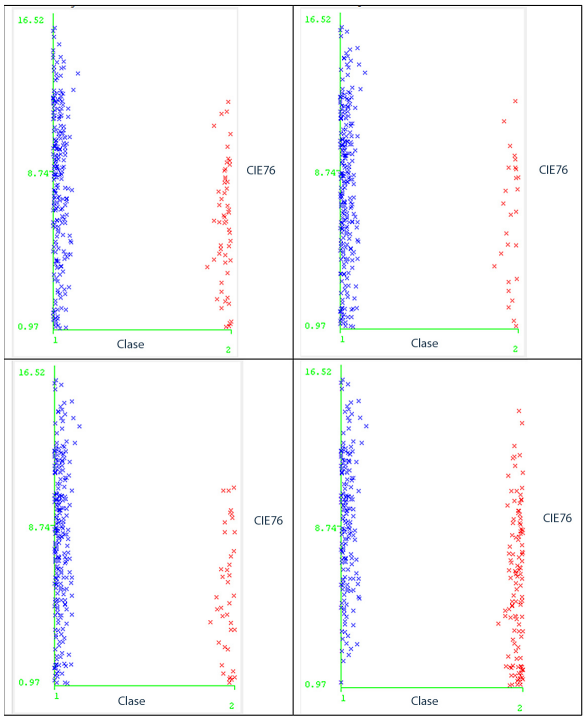
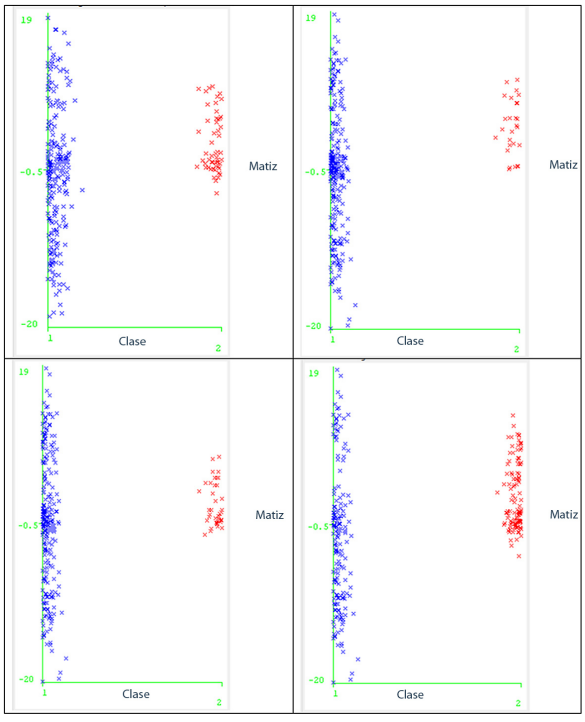


Figura 6. Distribución de imágenes inválidas (1) y válidas (2) en los rangos de valores de Matiz para los cuatro expertos (en orden del 1 al 4). Imagen 448



Rev. Esp. Doc. Cient., 39(2), abril-junio 2016, e128. ISSN-L: 0210-0614. doi: http://dx.doi.org/10.3989/redc.2016.2.1249

Figura 7. Distribución de imágenes inválidas (1) y válidas (2) en los rangos de valores de Saturación para los cuatro expertos (en orden del 1 al 4). Imagen 448

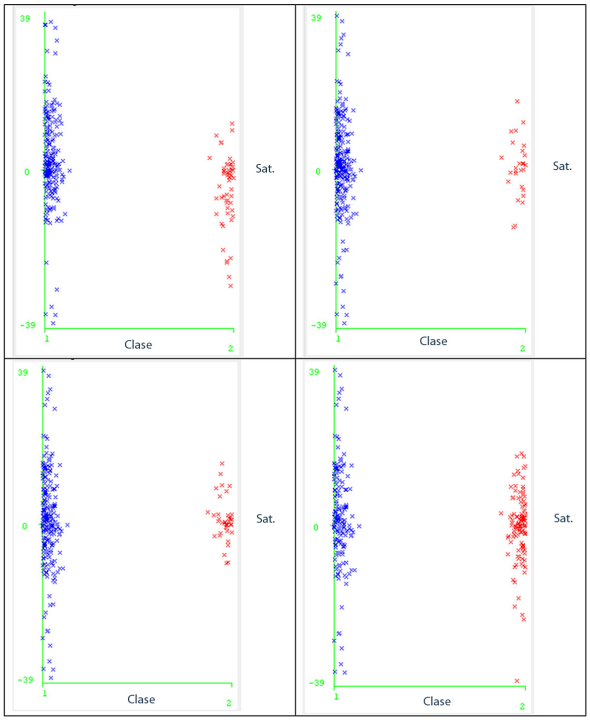
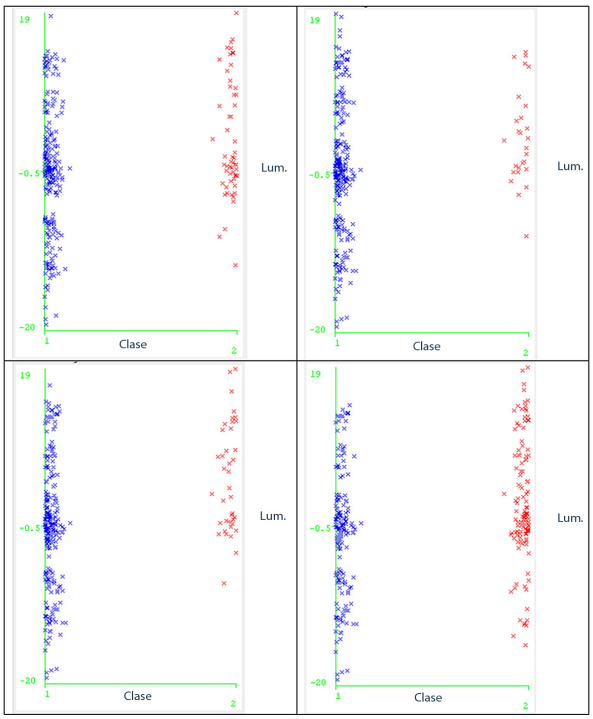


Figura 8. Distribución de imágenes inválidas (1) y válidas (2) en los rangos de valores de Luminosidad para los cuatro expertos (en orden del 1 al 4). Imagen 448

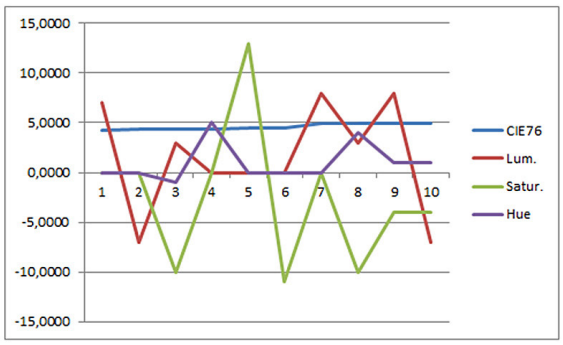


rangos es muy alto, lo que impide establecer rangos de aceptación fijos sobre una sola métrica o variable perceptual de color de forma aislada. Parece que los expertos varían su criterio para cada tipo de imagen, no estando siempre en el mismo valor para las tres imágenes el rango de aceptación de las diversas métricas. Por ello, podemos concluir que el motivo icónico de la imagen es determinante con respecto al grado de exigencia que aplica el experto y a la percepción de problemas de color y tonales. A la vista de los resultados habría que revisar la utilidad de los rangos de aceptación fijos en las métricas CIE76 y CIE00 que podemos encontrar en muchos sistemas de control de calidad de imágenes, pues los rangos de aceptación son mucho más amplios que los considerados comúnmente en los sistemas de control de calidad patrimoniales y admiten un alto porcentaje de imágenes no válidas, para cuyo descarte habría que considerar también el rendimiento en atributos perceptuales de color y sus interrelaciones, aspectos que no parecen estar suficientemente bien modelados en las métricas CIE que hemos empleado.

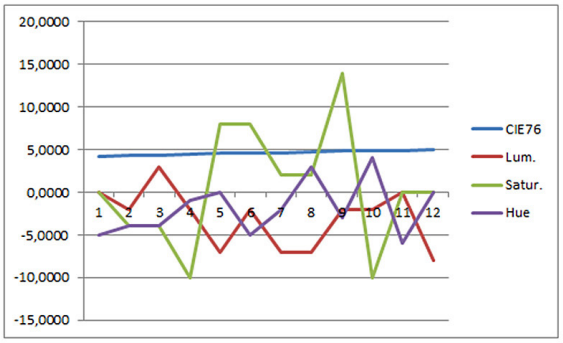
Para poder comprobar el grado de similitud en los patrones de correlación entre juicios de calidad y variaciones de valor en los parámetros analizados de los diferentes expertos, hemos estudiado con detenimiento qué ocurre en las zonas de solapamiento. Las zonas de solapamiento son los intervalos dentro de los valores de una variable donde se dan tanto imágenes válidas como inválidas. La finalidad de este análisis es poder llegar a determinar los factores que provocan la consideración de válida o inválida una imagen dentro de esas zonas por cada experto y si existe una pauta regular en la actuación de estos factores que nos ayude a conseguir un modelo. Hemos analizado el papel que juega la variabilidad en HSL para que las imágenes se consideren válidas o inválidas dentro de un mismo intervalo, considerando la métrica CIE76 y una de las imágenes, la 448. Para ello hemos representado los datos individuales de cada experto y de cada imagen, dividiendo los valores de la métrica CIE76 en intervalos, entre los rangos 1 y 8 (deltas 1 a 8), y estudiando cómo se comporta la variación en las variables HSL. En el eje X hemos representado el número de orden de las imágenes, y en el Y el valor de las variables HSL y CIE76. Las imágenes han sido ordenadas de menor a mayor por su valor CIE76. Para simplificar los resultados presentamos sólo los datos de la comparativa entre los expertos 4 y 1, la imagen 448 y los rangos CIE76 delta 4 (Figura 9), 6 (Figura 10) y 8 (Figura 11). Enfrentamos a izquierda y derecha los gráficos de las imágenes válidas e inválidas para facilitar la visualización de regularidades en los patrones de ambos tipos. Las líneas que representan las métricas y atributos perceptuales son de diferentes colores: CIE76, azul; Luminosidad, rojo; Saturación, verde; Matiz, violeta.

Si atendemos a estos gráficos, podemos observar cómo los patrones numéricos de las variables perceptuales de color HSL son muy diferentes den-

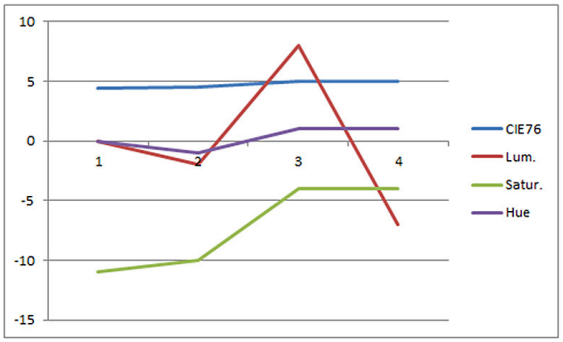
Figura 9. Rango delta 4



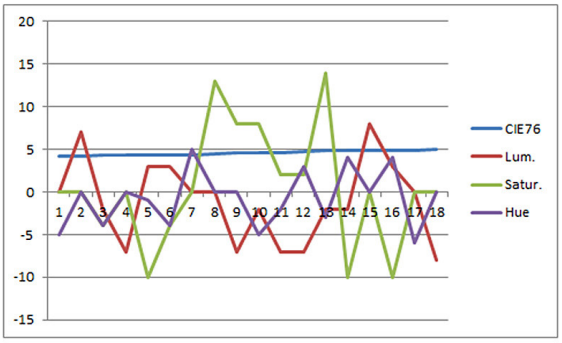
Experto 4. Delta 4. Imágenes válidas



Experto 4. Delta 4. Imágenes inválidas

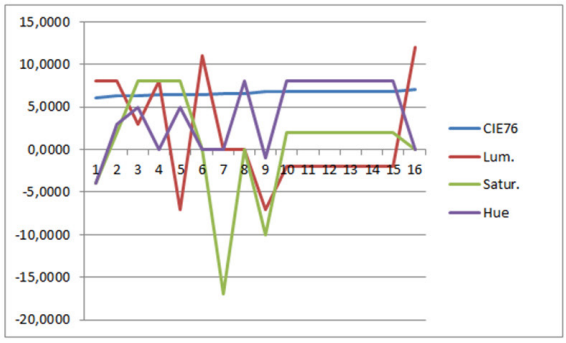


Experto 1. Delta 4. Imágenes válidas

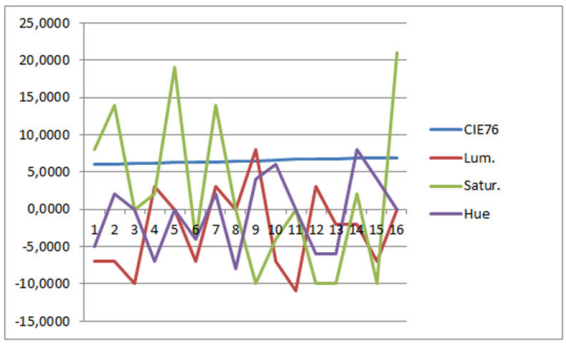


Experto 1. Delta 4. Imágenes inválidas

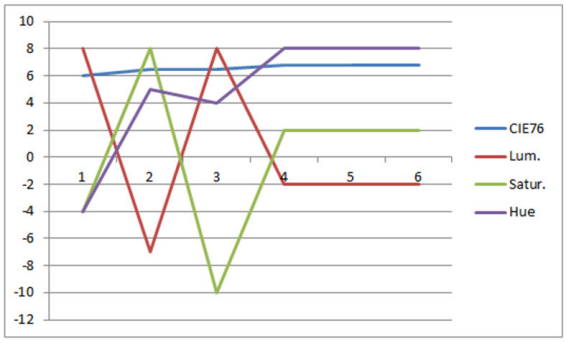
Figura 10. Rango delta 6



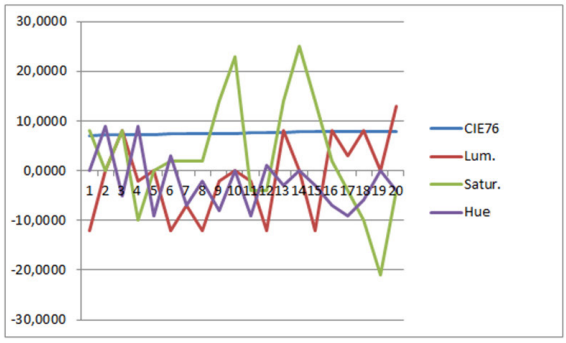
Experto 4. Delta 6. Imágenes válidas



Experto 4. Delta 6. Imágenes inválidas



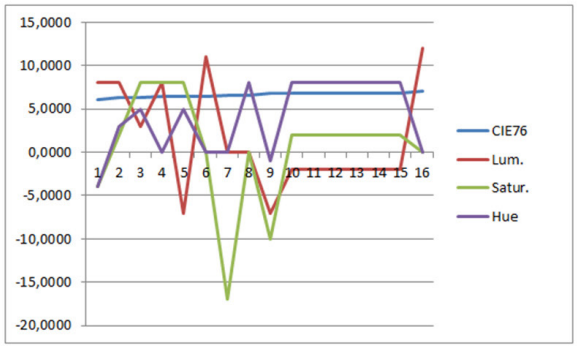
Experto 1. Delta 6. Imágenes válidas



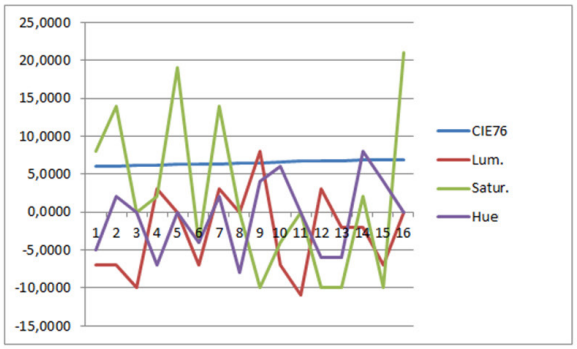
Experto 1. Delta 6. Imágenes inválidas

Rev. Esp. Doc. Cient., 39(2), abril-junio 2016, e128. ISSN-L: 0210-0614. doi: http://dx.doi.org/10.3989/redc.2016.2.1249

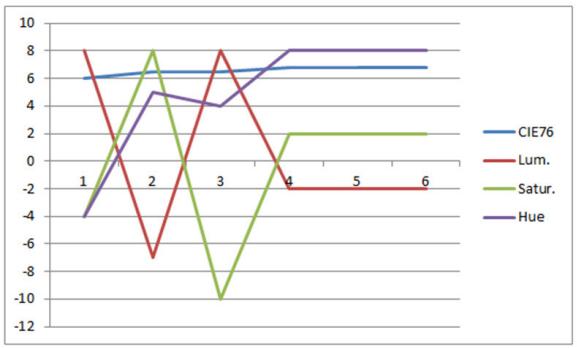
Figura 11. Rango delta 8



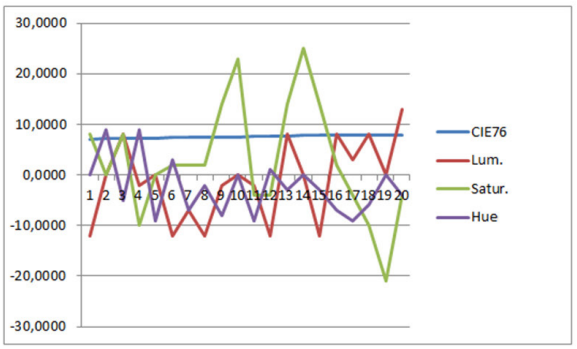
Experto 4. Delta 6. Imágenes válidas



Experto 4. Delta 6. Imágenes inválidas



Experto 1. Delta 6. Imágenes válidas



Experto 1. Delta 6. Imágenes inválidas

tro de un mismo rango de delta entre las imágenes aceptadas y las no aceptadas en los dos evaluadores; aunque también aparecen coincidencias, son muy escasas. Si asumimos el mismo porcentaje de inconsistencia que han tenido los expertos en las imágenes repetidas para el resto de imágenes del conjunto experimental, es lógico encontrar repetidos patrones de rechazo dentro de los patrones de aceptación y viceversa. Pero las diferencias encontradas en los patrones refuerzan la idea de que no es posible basar los modelos de control de calidad en unos rangos fijos de aceptación sobre métricas CIE Delta E 1976 o CIEDE 2000 sin considerar también comportamientos en las variables Matiz, Saturación y Luminosidad. Por ello, los modelos rígidos basados en rangos de métricas consideradas aisladas no deberían ser usados para conseguir una evaluación de calidad eficaz.

Sí que se puede confirmar la existencia de un modelo numérico similar en los valores HSL de las imágenes válidas y de las no válidas en los expertos 1 y 4, que se va haciendo prácticamente idéntico en las no válidas mientras sube el delta. Esta progresiva similitud se explica porque al haber mayor número de imágenes no válidas en los deltas más altos es más factible que aumente gradualmente la coincidencia de ambos expertos en las imágenes que han elegido como válidas y como inválidas.

3.3. Rendimiento de un algoritmo de aprendizaje automático para la obtención de un algoritmo visual

A través del algoritmo C4.5 hemos obtenido un conjunto de reglas que permite clasificar nuevos ejemplos de imágenes como válidos o inválidos de forma a como lo haría el experto humano a partir de cuyos datos de evaluación ha inferido las reglas. A modo de ejemplo, en la Figura 12 incluimos el árbol de decisión para el experto 1 y la imagen 550.

Para medir el grado de eficacia y eficiencia hemos utilizado diferentes indicadores. Entre ellos incluimos las tasas de precisión y de llamada. La primera expresa de entre las imágenes recuperadas por el sistema de reglas dentro en una clase, por ejemplo, la clase 1 (no válidas), la proporción de las que son correctas por corresponder a su clase y las que no. La tasa de llamada expresa la proporción de imágenes de una clase que han sido correctamente asignadas por el sistema de reglas con respecto a todas las imágenes correspondientes a esa clase. Para la imagen 448 véase la Tabla XI. Para la imagen 449 véase la tabla XII. Para la imagen 550 véase la tabla XIII.

Los porcentajes de acierto del sistema de reglas son siempre superiores al 85%, destacando la imagen 449 donde supera el 91,5%. Las tasas de precisión y llamada, excepto en la imagen 448 para

Figura 12. Árbol de decisión para el experto 1 y la imagen 550

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

Tabla XI. Rendimiento del sistema de reglas para imagen 448

Experto 1									
Total de imágenes	Correctamente clasificadas		Incorrectamente clasificadas		Precisión		Llamada		Número de reglas
471	Total	%	Total	%	Vál.	Inval.	Vál.	Inval.	33
	406	86.38	64	13.61	0.78	0.98	0.98	0.75	
Experto 4									
Total de imágenes	Correctamente clasificadas		Incorrectamente clasificadas		Precisión		Llamada		Número de reglas
728	Total	%	Total	%	Vál.	Inval.	Vál.	Inval.	40
	670	92.03	58	7.97	0.90	0.94	0.95	0.89	

Tabla XII. Rendimiento del sistema de reglas para imagen 449

Experto 1									
Total de imágenes	Correctamente clasificadas		Incorrectamente clasificadas		Precisión		Llamada		Número de reglas
525	Total	%	Total	%	Vál.	Inval.	Vál.	Inval.	14
	482	91.80	43	8.19	0.90	0.94	0.95	0.89	
Experto 4									
Total de imágenes	Correctamente clasificadas		Incorrectamente clasificadas		Precisión		Llamada		Número de reglas
515	Total	%	Total	%	Vál.	Inval.	Vál.	Inval.	13
	472	91.65	43	8.35	0.86	1	1	0.83	

Tabla XIII. Rendimiento del sistema de reglas para imagen 550

Experto 1									
Total de imágenes	Correctamente clasificadas		Incorrectamente clasificadas		Precisión		Llamada		Número de reglas
504	Total	%	Total	%	Vál.	Inval.	Vál.	Inval.	16
	470	93.44	33	6.56	0.89	1	1	0.87	
Experto 4									
Total de imágenes	Correctamente clasificadas		Incorrectamente clasificadas		Precisión		Llamada		Número de reglas
474	Total	%	Total	%	Vál.	Inval.	Vál.	Inval.	19
	436	91.98	38	8.02	0.87	0.99	0.99	0.86	

el experto 1, son superiores siempre al 0,83. Las tasas de inconsistencia que hemos comprobado que tienen todos los expertos impedirían obtener un sistema de reglas directamente del análisis de su comportamiento que tenga un rendimiento del 100%, pues el sistema de reglas modela en cierta medida esa inconsistencia al inferir las reglas directamente de los datos obtenidos de los propios expertos.

Podemos asumir que el resultado de la prueba de aprendizaje automático refuerza la conclusión obtenida en el epígrafe anterior sobre la existencia de patrones regulares en los juicios de calidad de los expertos, que esos patrones se basan en el análisis visual de propiedades perceptuales del color y que es posible generar un modelo que represente esos patrones regulares mediante el uso combinado de métricas y atributos perceptuales de color fácilmente computables, tales como CIE76, CIE00 o HSL. Por tanto, entendemos que es posible generar un modelo numérico, que con un reducido juego de variables, arroje una tasa de acierto relativamente alta si la comparamos con las tasas de error que hemos encontrado en las evaluaciones de los expertos humanos participantes en el experimento. La representación matemática de ese modelo conformaría un algoritmo visual. Para poder valorar la complejidad que podría llegar a tener un algoritmo visual de estas características basado en reglas hemos de analizar la complejidad de los árboles de decisión. Salvo en la primera imagen, los tamaños obtenidos son reducidos, pues el número de reglas oscilan entre 13 y 19. En la primera imagen lo hacen entre las 33 y 40. Por ello, el algoritmo visual sería realmente eficiente con la potencia informática a nuestra disposición actualmente.

Hemos aplicado a continuación el algoritmo C4.5 exclusivamente para las métricas CIE76 y CIE00 aisladas, sin considerar los atributos HSL, usando los datos del experto 4, el más consistente, y la imagen en la que el algoritmo ofrece mayor porcentaje de acierto, la 448. Los resultados obtenidos de inferir las reglas sólo con las métricas CIE aisladas no son aceptables, pues el porcentaje de acierto es muy bajo, 66,4% para CIE76 y 61,5% para CIE00, como era de esperar tras observar el alto grado de solapamiento que se produce entre los datos de las imágenes válidas y no válidas de acuerdo a estas métricas.

Hemos de reflexionar sobre la disparidad de resultados entre las reglas obtenidas para cada imagen y por cada experto. Esta disparidad implica que los criterios que se aplican para los juicios de valor varían según el motivo de la imagen, y según el experto. Ambos tipos de inconsistencia son un problema para los sistemas de evaluación de calidad basados en expertos humanos. Por ello, se hace preciso realizar estudios que aborden con mayor profundidad cómo influye el tipo de motivo de la imagen en la percepción de calidad, y cuáles son los factores que provocan la falta de consistencia entre evaluadores. Los métodos de análisis que hemos empleado para este estudio pueden ser empleados para la detección y análisis de este tipo de problemas.

4. CONCLUSIONES

Los sistemas de control de calidad de digitalizaciones patrimoniales deben considerar el rendimiento de los parámetros de medida de calidad no sólo a nivel físico, sino también perceptual global, modelando en la medida de lo posible las complejas interacciones que a este nivel se dan entre los atributos de calidad de la imagen. Un modelo perceptual implica un conocimiento que debe ser obtenido mediante la experimentación con expertos humanos en calidad suficientemente formados en los objetivos de los proyectos. Estos experimentos chocan con el problema de la inconsistencia inter e intra-evaluadores, que debe ser medida previamente.

Concluimos que no puede hablarse de rangos de aceptación continuos para las métricas consideradas habitualmente en los sistemas de calidad en color y en el uso de estas métricas de forma aislada, por lo que se ha de indagar en modelos más complejos. En este estudio hemos tratado de obtener un modelo basado en un sistema de reglas con alto rendimiento para el caso considerado en el experimento presentado empleando las métricas CIE76 y CIE00 junto a los atributos perceptuales del color HSL. La detección de patrones de valores de estos atributos regulares en las zonas de solapamiento entre imágenes consideradas como válidas e inválidas por los expertos, nos ha conducido a considerar que esta combinación de atributos y métricas puede ser idónea para medir objetivamente la apreciación subjetiva de proximidad perceptual con un grado de acierto relativamente alto, que estará siempre limitado por los errores que comenten los expertos evaluadores humanos en su trabajo de evaluación.

Los resultados obtenidos tras la aplicación del algoritmo de aprendizaje automático C4.5 indican que es posible emular el proceso de valoración del experto con porcentajes de eficacia superiores al 85%. El porcentaje de error cometido por los expertos ha sido estimado entre un 10,87% y un 20%, por lo que podemos equiparar sus tasas de acierto con las del sistema obtenido creado. Dada la variabilidad de criterio detectada inter e intra-evaluadores, no puede generalizarse un único modelo para todo el conjunto de evaluadores, aunque es presumible que tras un período de formación suficientemente amplio y de puesta en común de resultados sea posible mejorar esa inconsistencia como para poder generar un único modelo altamente eficaz.

5. REFERENCIAS

Charrier, C.; Lézoray, O.; Lebrun, G. (2012). Machine learning to design full-reference image quality assessment algorithm. *Signal Processing: Image Communication*, 27, 209-219. http://dx.doi.org/10.1016/j.image.2012.01.002

Dormolen, H. (2012). *Metamorfoze Preservation Imaging Guidelines. Test Version 1.0, January 2012*. https://www.metamorfoze.nl/sites/metamorfoze.nl/files/publicatie_documenten/Metamorfoze_Preservation_Imaging_Guidelines_1.0.pdf [29/12/2014].

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

Engeldrum, P. G. (2004). A Theory of Image Quality: The Image Quality Circle. *Journal of Imaging Science and Technology*, vol. 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*. http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf [20/04/2014].

Fairchild M. D. (2004). *Color Appearance Models: CIECAM02 and Beyond*. IS&T/SID 12th Color Imaging Conference. Tutorial T1A, 11/9/04. http://www.cis.rit.edu/fairchild/PDFs/AppearanceLec.pdf [20/05/2014].

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*. (2ª 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, Hsien-Che. (2005). *Introduction to Color Imaging Science*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511614392

Luo, M. R.; Cui, G.; Rigg, B. (2001). The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research & Application*, vol. 26 (5), 340–350. http://dx.doi.org/10.1002/col.1049

Martens, J.B. (2002). Multidimensional modeling of image quality. *Proceedings of the IEEE*, vol. 90 (1), 133-153. http://dx.doi.org/10.1109/5.982411

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). *Actas de las Primeras Jornadas sobre Imagen, Cultura y Tecnología*, pp. 99-120. Getafe, España: Universidad Carlos III de Madrid.

Nationaal Archief (2010). *Digitisation of photographic materials. Guidelines. September 2010*. http://www.nationaalarchief.nl/sites/default/files/docs/guidelines_digitisation_photographic_materials.pdf [19/11/2014].

Pellacini, F.; Ferwerda, J.A.; Greenberg, D.P. (2000). Toward a psychophysically-based light reflection model for image synthesis. *Proc. ACM SIGGRAPH 2000*, pp. 55-64. http://dx.doi.org/10.1145/344779.344812

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*. http://www.archives.gov/preservation/technical/guidelines.pdf [14/11/2014].

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

Robledano Arillo, Jesús (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*, vol. 34 (3), 357-384. http://dx.doi.org/10.3989/redc.2011.3.816

Robledano Arillo, J. (2011b). Twenty-five years of digital conversion. Current situation. En: Internacional Conference. *Thirty Years of Photographic Conservation Science.* Logroño (La Rioja). Spain. June, 2011.

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

Still Image Working Group (2010). *GAP Analysis. Updated 01/12/2010*. http://www.digitizationguidelines.gov/guidelines/Gap_Analysis.pdf [19/11/2014].

Williams, D. (2002). Image quality metrics. *RLG Diginews*, vol. 4 (4). http://www.worldcat.org/arcviewer/1/OCC/2007/08/08/0000070511/viewer/file1806.html [19/12/2014].

Williams, D. (2003). Debunking of Specsmanship: Progress on ISO/TC42 Standards for Digital Capture Imaging Performance. *IS&T's 2003 PICS Conference*, pp. 77-81.

Williams, D. (2010). *Imaging Science for Archivists*. http://www.digitizationguidelines.gov/guidelines/Digital_Imaging_Science.ppt [28/11/2014].

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

Zhou Wang; Bovik, A.C.; Ligang Lu (2002). Why is image quality assessment so difficult? *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on* (Volume: 4), p. IV-3313 - IV-3316. http://dx.doi.org/10.1109/icassp.2002.5745362