Automatizing chromatic quality assessment for cultural heritage image digitization

Ana Granados; Valentín Moreno-Pelayo; Jesús Robledano-Arillo

How to cite this article:

Granados, Ana; **Moreno-Pelayo, Valentín**; **Robledano-Arillo, Jesús** (2019). "Automatizing chromatic quality assessment for cultural heritage image digitization". *El profesional de la información*, v. 28, n. 3, e280305.

https://doi.org/10.3145/epi.2019.may.05

Manuscript received on August, 02th 2018
Accepted on February, 18th 2019

Ana Granados ✉
https://orcid.org/0000-0003-0158-7969

CES Felipe II, Universidad Complutense de Madrid
Capitán, 39. 28300 Aranjuez (Madrid), Spain
ana.granados@ajz.ucm.es

Valentín Moreno-Pelayo
https://orcid.org/0000-0002-8731-7443

Universidad Carlos III de Madrid
Computer Science and Engineering Department
Avda. Universidad, 30. 28911 Leganés (Madrid), Spain.
vmpelayo@inf.uc3m.es

Jesús Robledano-Arillo
https://orcid.org/0000-0002-4090-8684

Universidad Carlos III de Madrid
Library and Information Science Department
Madrid, 126. 28903 Getafe (Madrid), Spain.
jroble@bib.uc3m.es

Abstract

In the context of digitization of photographs and other documents with graphical value, cultural heritage organizations need to give a guarantee that the stored digital image is a faithful representation of the physical image both at the physical level and the perceptual level. On the physical level, image quality can be measured objectively in a simple way by applying certain physical attributes to the image, as well as by measuring how distorting images affects the performance of the attributes. However, on the perceptual level, image quality should correspond to the perception that a human expert would experience when observing the physical image under certain determined and controlled conditions. In this paper we address the problem of image quality assessment (IQA) in the context of cultural heritage digitization by applying machine learning (ML). In particular, we explore the possibility of creating a decision tree that mimics the response of an expert on cultural heritage when observing cultural heritage images.

Keywords

Photography collections; Cultural heritage digitization; Image quality assessment (IQA); Color; Machine learning (ML); Decision trees; Automatic classification; Learning models; Algorithms.

1. Introduction, problem statement and research questions

Cultural heritage images have been subject to digitization activities due to many reasons. For example, many cultural heritage organizations, such as libraries, archives or museums, want to digitally archive their print media collections, whereas others want to make the content available for end-users on a grand scale (**Liu** *et al.*, 2012). In this regard, digitizing cultural heritage allows preservation of the original cultural artifacts while enabling user interactivity through advanced multimedia presentation and wider distribution channels for greater educational accessibility (**Zhao**; **Campisi**; **Kundur**, 2004). Regardless of the reasons that make such organizations digitize cultural heritage images, assessing the quality of the digitized images has become crucial in order to storage faithful representations of the originals.

Image quality can be assessed subjectively (with human intervention) or objectively (without human intervention). Subjective evaluation is the most reliable method of quantifying visual image quality for applications in which images are ultimately to be viewed by human beings (**Wang** *et al.*, 2004). However, in practice, this heavy human intervention is time-consuming and expensive. This is the reason why objective evaluation has been widely applied to the field of image quality assessment (IQA). Objective image quality evaluation can be classified into two broad types: signal fidelity measures and perceptual visual quality metrics (**Zhang** *et al.*, 2011). Several signal fidelity measures have been used in the literature to assess image quality. Signal-to-noise ratio (SNR), peak signal-to-noise ratio (PSNR) and mean-squared error (MSE) are good examples of well-known signal fidelity measures that have been widely applied. Although these kinds of measures have been successfully applied, they present an issue: they do not align well with human visual perception (**Zhang** *et al.*, 2011). With the aim of addressing this issue, perceptual quality metrics have been developed, see for example (**Aydin** *et al.*, 2008; **Cadík** *et al.*, 2012; **Kopf** *et al.*, 2012; **Oeztireli**; **Gross**, 2015).

> Assessing the quality of the digitized images has become crucial in order to storage faithful representations

In the context of digitization of heritage photographs and other documents with graphical value, cultural organizations need to give a guarantee that the stored digital image is a faithful representation of the physical image both at the physical level and the perceptual level. Only in this way can stored images be used for the functions of custody, conservation, reproduction, analysis, study and dissemination which they are meant to support, within certain ethical criteria that do not approve any change in the plastic characteristics or reinterpretation of the iconic and plastic messages (*Fadgi*, 2010; **Frey**; **Reilly**, 2006; **Robledano-Arillo**, 2016; **Robledano-Arillo**; **Navarro-Bonilla**, 2017). In this sense, on the physical level, digital image quality can be measured objectively in a simple way by applying signal fidelity measures on a variety of physical attributes, for example, those established in ISO 19293 and 19264 standards (*ISO*, 2017a; *ISO*, 2017b). However, on the perceptual level, image quality should correspond to the perception that a human expert would experience when observing the physical image under certain determined and controlled conditions.

Given the ease of computing attributes and distortions of a physical nature, one important line of research into the development of IQA systems has focused on connecting the physical and perceptual levels, such that the overall quality of an image at the perceptual level can be automatically derived through the use of easily computable physical measures (**Engeldrum**, 1995; 2004). This is the focus of the research on perceptual quality metrics mentioned above. But nowadays we do not have perceptual metrics that can adequately replace the work of a human being in work contexts where high precision quality measures are required, such as cultural heritage digitization and dissemination.

> One important line of research into the development of IQA systems has focused on connecting the physical and perceptual levels, such that the overall quality of an image at the perceptual level can be automatically derived through the use of easily computable physical measures

In the field of cultural heritage, the main attempts to do that have been essentially focused on using a limited set of physical attributes for which certain previously determined value acceptance ranges are established. In other words, a digitized image has been considered to have a good quality if several attributes are within several ranges of values, as proposed in (**Williams**; **Longman**, 2003; **Puglia**; **Reed**; **Rhodes**, 2004; *Fadgi*, 2010; **Van-Dormolen**, 2012; *Nationaal Archief of The Netherlands*, 2010). The two most recent ISO standards on quality assessment in cultural heritage digitalization are based on this approach (*ISO*, 2017a; 2017b). This constitutes an important drawback because, as said above, in this context a high level of color fidelity is required. This implies that the human expert perception should be considered.

Due to this, in the field of cultural heritage digitization, the presence of experts on cultural heritage in the quality assessment process is needed. This dependence on cultural heritage experts makes the task of assessing image quality really expensive. Therefore, developing methods that enable automatization of the IQA process becomes essential.

In this regard, machine learning (ML) can be very helpful because ML methods allow the IQA task to be addressed from a different perspective. Thus, instead of modeling the human visual system (HVS), ML methods can be used to mimic the HVS response to quality losses (**Gastaldo** *et al.*, 2005). In particular, in the field of cultural heritage digitization, the goal would be to mimic the HVS response that an expert on cultural heritage would experience when observing the physical image.

In this study we address the problem of IQA in the context of cultural heritage digitization by applying ML. In particular, we create a decision tree that mimics the response of an expert on cultural heritage when observing several images. The paper is structured as follows. Section 3. Methodology describes how the dataset has been created, and how the decision tree is created from the knowledge extracted thanks to the experts on cultural heritage. Section 4. Experimental results shows and analyzes the obtained results. Section 5. Discussion provides a discussion of the experimental results. Finally, Section 6. Conclusions summarizes the conclusions of our work.

> Developing methods that enable automatization of the image quality assessment (IQA) process becomes essential

2. Literature review

Perceptual quality metrics can be classified according to the availability of a reference image as full reference (FR), reduced reference (RR) and no reference (NR). The first approach has given rise to well-known metrics such as the structural similarity index (SSIM) (**Wang** *et al.*, 2004), the multi-scale extension of SSIM (MS-SSIM) (**Wang**; **Simoncelli**; **Bovik**, 2003), the visual information fidelity (VIF) metric (**Sheikh**; **Bovik**, 2006), or the HDR-VDP-2 (**Mantiuk** *et al*., 2011), among others. The second approach has been applied in works such as (**Kusuma**; **Zepernick**, 2003), where the authors present the hybrid image quality metric (HIQM); (**Wang**; **Simoncelli**, 2005), where the authors propose an RR method based on a natural image statistic model in the wavelet transform domain; (**Li**; **Wang**, 2009), where an RR algorithm based on a divisive normalization image representation is presented, or (**Ma** *et al.*, 2011), where an RR method based on statistical modeling of the discrete cosine transform (DCT) coefficient distributions is presented. The third approach has been applied in many works as well. For example, in **Sheikh**, **Bovik** & **Cormack** (2005), the authors propose to use natural scene statistics (NSS) to blindly measure the quality of images compressed by JPEG2000 (or any other wavelet based) image coder. NSS has been applied also in **Brandão** & **Queluz** (2008), where a generic NSS-based NR algorithm that operates in the discrete cosine transform domain is proposed for JPEG or MPEG; or in **Mittal**, **Moorthy** & **Bovik** (2012), where an NSS-based distortion-generic NR model that operates in the spatial domain is proposed.

Numerous works have applied ML techniques to assess image quality from several approaches. For example, in **Chang** *et al.* (2016) convolutional networks, random forest classifiers and support vector machines have been used to automatically assess the quality of series of photos taken of the same scene in order to provide an automatic photo triage method. In **Gastaldo** *et al.* (2005), a circular back-propagation neural network is proposed for evaluating the effects of image enhancement filters, using general pixel-based image features. In **Li**, **Bovik** & **Wu** (2011), a general regression neural network is used to approximate the functional relationship between image features and subjective mean opinion scores. In **Suresh**, **Babu** & **Kim** (2009), an extreme learning machine classifier, which is a specific type of neural network, is applied to assess the problem of IQA from a classification problem perspective. In **Narwaria**, **Lin** & **Cetin** (2012), support vector regression (SVR) is used to map the high dimensional feature vector into a perceptual quality score. In **Liu**, **Lin** & **Kuo** (2013), a multi-method fusion (MMF) approach based on SVR is applied to IQA and two MMF-based quality indices are proposed. In **Gao** *et al.* (2013), image quality is estimated using a multiple kernel learning algorithm. In **De** & **Sil** (2011), a fuzzy relational classifier that assesses the quality of images is proposed. There even exist papers that evaluate and compare other works that apply ML techniques to IQA (**Gastaldo**; **Zunino**; **Redi**, 2013; **Charrier**; **Lézoray**; **Lebrun**, 2012).

3. Methodology

In this section, we describe how the dataset has been created, how the experiments have been set up, and how the decision tree has been created from the knowledge extracted thanks to the experts on cultural heritage.

3.1. Dataset creation

In the field of cultural heritage digitization, image quality is assessed by an expert on cultural heritage who compares the physical image and the digital image. This implies that the quality of the image cannot be assessed by any human, but by an expert on cultural heritage. This is the reason why, in our work, we do not use common image databases such as *LIVE* (**Sheikh**; **Sabir**; **Bovik**, 2006), *CSIQ* (**Larson**; **Chandler**, 2010) or *TID2013* (**Ponomarenko** *et al.*, 2015), which are based on mean opinion scores (MOS), but we create our own database.

Our dataset has been created by carrying out a psychometric experiment in which several experts on cultural heritage have evaluated a set of degraded images comparing them to the corresponding physical images. The physical originals have been compared to the corresponding on-screen digital images according to the standards for carrying out quality evaluation *ISO 12646:2015*, *ISO 3664:2009* and *ISO 20462-3:2012* (*ISO*, 2015; 2009; 2012).

The creation of our dataset comprises three phases:

- Digitization of photographic images.
- Creation of degraded images.
- Assessment of image quality by the experts on cultural heritage.

Three photographic images on paper that are representative of the type of documents found in many photography collections have been used to create our dataset (Fig. 1, Fig. 2 and Fig. 3). The digitized masters of the original images have been created using a digital single-lens reflex camera. In order to obtain images with high fidelity in color and contrast at the colorimetric and densitometric level, color management through customized ICC profiles has been applied. The colorimetric fidelity data obtained in the patches of the *ColorChecker*® color chart digitized together with the physical originals can be seen in Table 1.

Table 1. Colorimetric fidelity of master images, showing deltas from each master image with respect to the colorimetric values of *ColorChecker*® patches

	DeltaE 1976	*Ciede 00*
Master 1	0.97	0.66
Master 2	1.12	0.80
Master 3	0.87	0.69

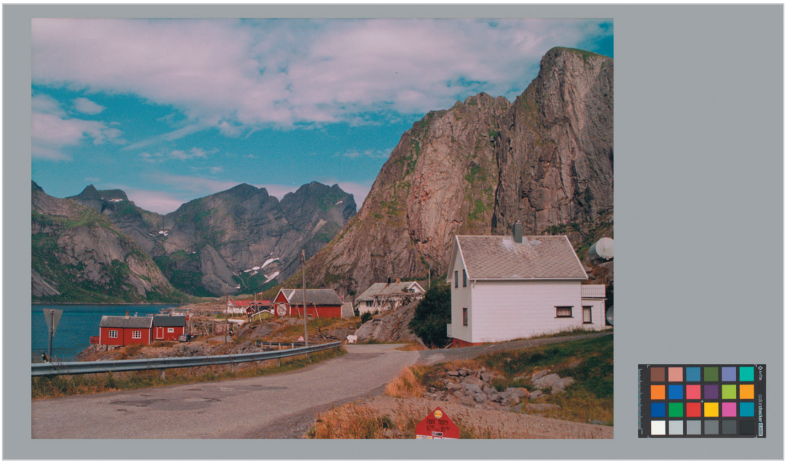

Figure 1. Mountain and water landscape. Matte color photographic print. Late 20th century.

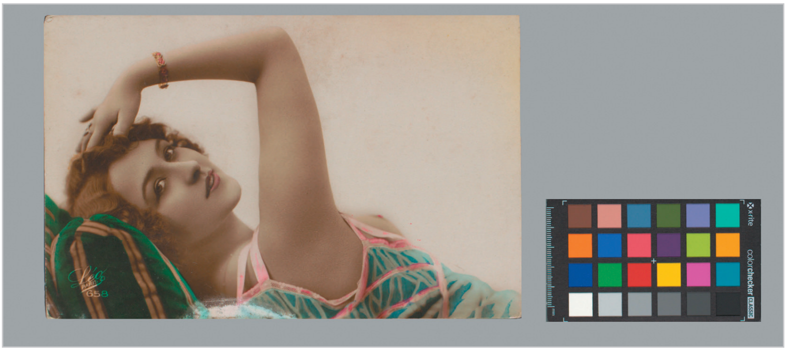

Figure 2. Human portrait. Monochromatic photographic print, hand-colored in ink. Early 20th century.

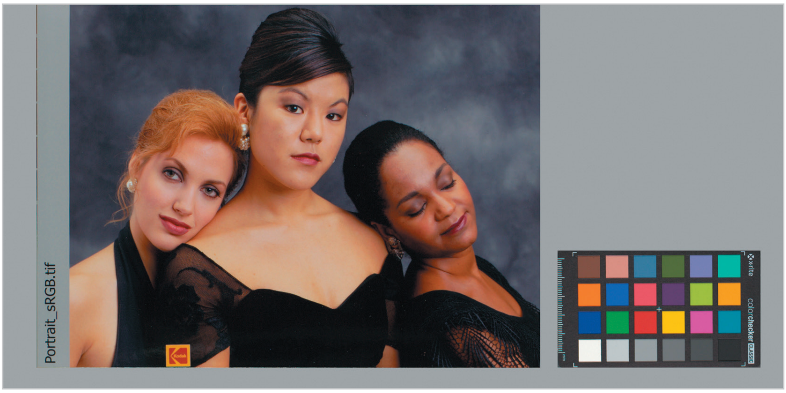

Figure 3. Human portrait. Glossy color photographic print. Early 21st century.

Based on the masters, a series of 300 degraded images per physical original have been created by editing their HSL perceptual values: Hue (H), Saturation (S) and Lightness (L). This has created a degradation sequence that contemplated a sufficiently broad scale of perceptible changes in these three color-description variables. To do this, the images have been converted to the HSL color space and progressively degraded in these three variables:

- Hue: from -20 to +19 (on a scale ranging from -100 to +100).
- Saturation: from -39 to +39 (on a scale ranging from -100 to +100).
- Lightness: from -20 to +20 for (on a scale ranging from -100 to 99).

We have only focused on color problems that affect the image globally (i.e. in all its surface) in order to simplify the experimentation because local problems of the spatial inhomogeneity type (light fall-off or vignetting, for example) are usually detected and corrected in a professional work environment prior to or after capture.

In addition, with the aim of analyzing the reliability of the human experts' judgements, several repeated images have been included in each series of images to be evaluated. Said repeated images have been used to analyze the consistency of the experts' judgements. That is, we have analyzed whether the response of the human expert for a given image is always the same.

Once all the images have been created, several data for each image have been recorded automatically by applying various comparison metrics for color and image difference between the digital masters and their degradations. The color differences between the digital images and the degraded digital images have been calculated using all the patches from the *ColorChecker*®.

As mentioned previously, the visual evaluation performed by the experts has been accomplished according to the standards for carrying out quality evaluation *ISO 12646:2015*, *ISO 3664:2009* and *ISO 20462-3:2012*. Four experts on cultural heritage have been selected who satisfied the condition of being professionals with extensive experience in the sectors of professional photography and graphic arts. The visual evaluation carried out by the experts consisted of comparing the physical originals to the corresponding on-screen digital images, both the expert and the physical original being inside a viewing booth. In this process, all the elements that comprise the viewing flow have been totally controlled. These, in addition to the digital image, include the following:

- the calibration and ICC profile of the monitor;
- the conversion from the color space of the image to the color space of the monitor made by the operating system's color management system (CMS);
- the quality of the monitor and of the conditions of its viewing environment;
- the quality of the viewing booth for the physical originals and its viewing conditions.

Table 2 shows calibration conditions according to *ISO 12646* and Table 3 shows the adjustment and calibration data for the experiments viewing environment.

Table 2. Viewing conditions according to *ISO 12646*

Monitor white point and display booth color temperature	CIE D50 or 5.000K
Monitor white point luminance	Between 80 cd/m^2 and 160 cd/m^2
Ambient light white point	CIE D50
Environment luminance	32 lux
Display booth illuminance	500 lux ± 125
Environment	Neutral color with a maximum reflection of 60% of the light
Monitor gamma value	Between 1.8 y 2.4
Color temperature difference between ambient light and light from the viewing booth	± 200 K
Luminance difference between ambient light and monitor white point	Less than 1/10
CRI booth display	90%
Contrast ratio between the monitor white and black point	Above 100:1

Table 3. Experiment viewing conditions

Display booth color temperature	Between 5130K and 5160K
Monitor white point color temperature	5222K
Monitor white point luminance	113.04 cd/m^2. Around 450 lux
Gamma monitor in RGB channels	2.2
RGB monitor chromaticity	0.656,0.327; 0.212,0.684; 0.150,0.072
Light intensity in display booth	Around 600 lux
Ambient light color temperature	4504 K
Ambient light intensity	55.6 luxs
CRI booth display	84.6 %
Contrast ratio between the monitor white and black point	Above 410:1

The screen interface used to show the digital images has been designed using the *Adobe Bridge* program so that the screen only showed the image being evaluated and a narrow band along the left side of the screen that allowed the evaluators to navigate through the images. The physical image and the *ColorChecker*® used to create the digital masters have been placed in the booth in a position that was very similar to that in the test images. The intensity of the grey background color of the screen was made to coincide with that of the booth.

Based on the quality detected, the experts were able to assign a score to each image based on a scale with three values:

- The image would not pass a professional quality control measuring the proximity in the appearance of color and contrast between the image on the screen and the image on paper.
- The image would pass the quality control with a rigorous criterion.
- The image would pass the quality control with a less rigorous criterion.

The purpose of our study is to explore the possibility of generating a system of rules that emulates a human expert with a high percentage of success. For this reason, the best of the evaluators is chosen, based on the highest consistency of his evaluation judgments.

3.2. Experimental setup

Given that in the field of cultural heritage, automated digitization quality systems have been based essentially on using a limited set of physical attributes for which certain previously determined value-acceptance-ranges are established, in previous work, we analyzed the ranges of acceptance of several physical attributes for each expert on cultural heritage (**Robledano-Arillo**; **Moreno-Pelayo**; **Pereira-Uzal**, 2016). In said analysis, we observed that these ranges do not coincide at 100% for all of the experts. This prevented us from creating a decision tree based on the quality judgments given by all the experts. That is the reason why, in this work, we only use the judgements given by one of the evaluators. Specifically, the best of the evaluators according to the highest consistency of his results.

In the interest of simplifying our analysis, instead of allowing an image to have three different scores (1. not valid, 2. valid with a rigorous criterion, 3. valid with a less rigorous criterion), we have combined values 2 and 3 to work only with two quality classes: invalid image and valid image.

Out of the 903 images evaluated by the experts on cultural heritage, the best evaluator classified 223 images as "valid images'' and 680 images as "invalid images''. Thus, the number of valid and invalid images is unbalanced. This fact can affect the creation of the decision tree; therefore, instead of creating a decision tree from the 903 images, we have created three datasets, so that in each dataset we could have approximately 50% of valid images and 50% of invalid images. This is the number of valid/invalid image in each dataset:

- Dataset 1: 223 valid images + 227 invalid images
- Dataset 2: 223 valid images + 226 invalid images
- Dataset 3: 223 valid images + 227 invalid images

It has to be pointed out that the invalid images have been randomly sorted, so there is no bias in their later selection.

3.3. Decision-tree creation

As mentioned previously, ML techniques have been applied to assess image quality from several approaches. In this sense, approaches like neural networks (**Gastaldo** *et al.*, 2005; **Li**; **Bovik**; **Wu**, 2011; **Suresh**; **Babu**; **Kim**, 2009), support vector regression (**Narwaria**; **Lin**; **Cetin**, 2012), multiple kernels (**Gao** *et al.*, 2013) or fuzzy classifiers (**De**; **Sil**, 2011) have been used in the literature. In this work, the well-known C4.5 algorithm created by Ross Quinlan (**Quinlan**, 1992) has been used to generate a decision tree that classifies an image as valid or invalid. In terms of implementation, we have used the *Waikato Environment for Knowledge Analysis*, best known as *Weka* (**Hall** *et al.*, 2009), to create the decision tree.

Different physical attributes of the images could be used to create the decision tree. In order to analyze what selection of physical attributes provides a higher success rate, we have created several decision trees from each dataset. The attributes used in each test are the following ones:

- CIE76: Average of the color differences between the digitized physical image and real color from the *ColorChecker*® card.
- CIE76 + PSNR + SSIM: CIE76 measure plus the well-known peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM).
- CIE76 + HSL: CIE76 measure plus the HSL perceptual values: hue (H), saturation (S) and lightness (L).
- CIE76 + HSL + patches: CIE76 measure plus the HSL perceptual values plus the 24 color differences between the digitized physical image and the real color from each of the 24 patches from the *ColorChecker*® card.

Using each of these four sets of attributes, we have applied three algorithms to create our classifier. Thus, as well as using the C4.5 algorithm, we have applied ensemble learning in order to study different classification approaches. Ensemble learning is the process of generating multiple classifiers and combining their results to give a classification result. In this work, we have used two ensemble learning strategies: bagging and boosting.

Among the ensemble-based algorithms that exist, bagging is one of earliest and simplest ones and its performance is quite good (**Breiman**, 1996). In bagging, the classifiers are obtained by using bootstrapped replicas of the training data. In other words, different training data subsets are created by randomly drawing a percentage of the data contained in the entire training dataset. After that, each subset is used to train a different classifier. The results given by each of the classifiers are combined by taking a simple majority vote of their decisions. That is, in our case, an image would be considered to be valid if the majority of the classifiers classified it as valid.

In boosting, an ensemble of classifiers is also created by resampling the data. However, in this case, resampling is strategically oriented to provide the best possible training data for each consecutive classifier. Essentially, three weak classifiers are created in each iteration of boosting. The first classifier created is trained with a random subset of the training data. The second classifier is trained with the most informative subset, given the first classifier. Specifically, the second classifier is trained with some training data in which only half of the data is correctly classified by the first classifier, and the other half is misclassified. The third classifier is trained with instances on which the first and second classifiers disagree. Then, the three classifiers are combined through a three-way majority vote.

4. Experimental results

The success rate obtained for each of the tests performed can be found in Table 4. Besides, the average of the success rates for all the datasets has been included in order to ease the comparison of each pair selection-of-attributes and decision-tree-creation-algorithm.

Table 4. Experimental results. Success rate (%) for each configuration.

	CIE76			CIE76 + PSNR + SSIM			CIE76 + HSL			CIE76 + HSL + patches		
	C4.5	**Bagging**	**Boosting**	**C4.5**	**Bagging**	**Boosting**	**C4.5**	**Bagging**	**Boosting**	**C4.5**	**Bagging**	**Boosting**
Dataset1	68.75	65.63	65.63	64.96	66.52	66.29	76.34	76.12	75.45	80.13	75.67	78.35
Dataset2	66.52	64.51	63.17	70.54	71.88	69.42	81.03	79.69	79.46	79.91	92.19	99.11
Dataset3	71.74	69.76	68.65	70.86	72.41	72.19	63.13	60.49	63.80	84.77	89.40	86.76
Mean	69.00	66.63	65.82	68.79	70.27	69.30	73.50	72.10	72.90	81.60	85.75	88.07

Analyzing the results shown in Table 4, it can be observed that regardless of the learning algorithm used to create the classifier, the best results are always obtained when the physical attributes selected as input to the learning algorithm are CIE76+HSL+patches. That is, the best option is to use the following data as input to the algorithm:

- CIE76: Average of the 24 color differences between the digitized physical image and real color from the *ColorChecker*® card.
- HSL perceptual values: hue (H), saturation (S) and lightness (L).
- Patches: the 24 color differences between the digitized physical image and the real color from each of the 24 patches from the *ColorChecker*® card.

In order to ease the comparison of all the experimental results, a figure that summarizes the results obtained has been created. This figure (Fig. 4) depicts the success rates averaged across all the datasets, for each learning algorithm and each selection of attributes.

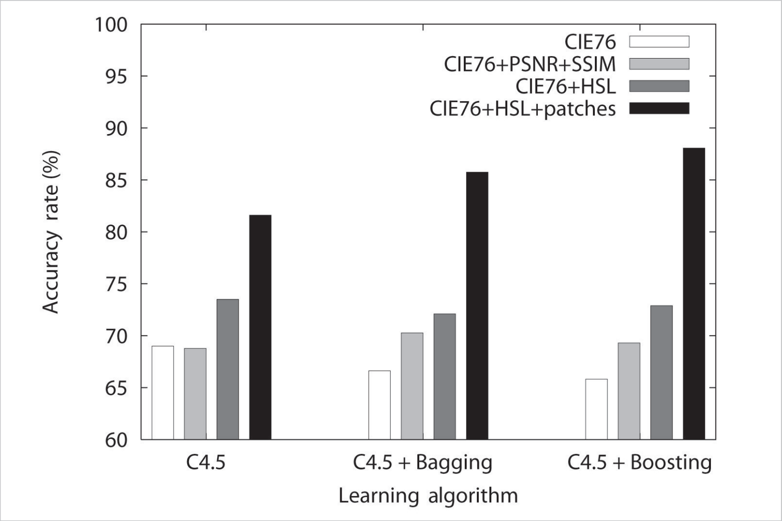

Fig. 4. Accuracy rates for each learning algorithm. Comparative of the success rates obtained for each learning algorithm and each selection of attributes.

Analyzing Fig. 4, we can observe that regardless of the learning algorithm used, the worst success rates are obtained when the CIE76 is used as input attributes to the algorithm. Obviously, metrics such as CIE76 or CIE00 should give the worst performances, since they average color differences of several chromatic attributes. However, we have evaluated these metrics in isolation to contrast the low perceptual value of automated color fidelity evaluations based solely on these kinds of metrics.

The success rates obtained using CIE76 as input to the algorithm are slightly improved when the attributes used as input are CIE76+PSNR+SSIM. This confirms our concerns about SSIM, which is made for structural deformation due to, for example, compression algorithms, and not chromatic deviations that maintain image structure substantially.

The success rates obtained using CIE76+PSNR+SSIM are lightly improved when the attributes used as input are CIE76+HSL, and finally, the best success rates are obtained when the attributes used as input to the learning algorithm are CIE76+HSL+patches. It has to be highlighted that using the patches improves the results because we are considering separately all the color patches without averaging.

It can be observed that the best results are obtained when the decision tree is created using CIE76, Hue, Saturation, Lightness and patches as input variables, regardless of the algorithm used to create the decision tree.

Therefore, the main conclusion that can be derived from the experimental results obtained is that when the input attributes are CIE76+HSL+patches, the best success rate is achieved when the classifier is created combining the well-known C4.5 algorithm and the ensemble learning strategy boosting.

It has to be highlighted that the best success rate obtained is close to the success rate of the expert. This can be due to the intrinsic characteristics of rule induction algorithms. Thus, these kinds of algorithms are robust in the presence of noise (errors, omissions or lack of data) during the rule induction process (**Major**; **Mangano**, 1995). This makes us suppose that the models obtained in this work will generalize the expert judgements, the models not being affected by the possible inconsistences in the judgements of the experts. However, it has to be kept in mind that if there were inconsistencies in the expert judgements, these inconsistencies would be included not only in the learning phase but also in the evaluation tests. Therefore, the evaluation results would be negatively affected. In other words, the maximum success rate obtained for the learning algorithm would be close to the success rate of the expert, even when the model behavior was adjusted to the data. This might be the reason why the best success rate obtained in our work is close to the success rate of the expert.

5. Discussion

5.1. The need to use more complex perceptual color attributes for quality assessment in the field of heritage photographs digitization

The metrics based on vectorial distance (CIE76, PSNR and SSIM) have worse performance than the metrics that consider perceptual variables separately (HSL). Thus, it seems that modeling image quality according to metrics that give a summary of the performance of a complex perceptual attribute through one single number is not recommendable. Example of such metrics could be CIE76 or CIEDE2000 (**Luo**; **Cui**; **Rigg**, 2001) for color, PSNR for noise, or SSIM for image structure. The success rates obtained when such metrics are used as input for the learning algorithm are poor. Thus, we could say that we should not express the quality distances in a multidimensional space by only one number, but by as many numbers as dimensions the space has.

When perceptual color attributes are used as independent attributes (e.g. color is decomposed in its HSL perceptual values), better success rates are obtained. In fact, the best results are obtained when the input attributes to the learning algorithm are CIE76+HSL+patches. That is, the inputs are the average of the 24 color differences between the digitized physical image and real color from the *ColorChecker*® card, the HSL perceptual values (hue, saturation and lightness), and the 24 color differences between the digitized physical image and the real color from each of the 24 patches from the *ColorChecker*® card.

5.2. Study limitations

In order to achieve a good performance applying ML, we had to discard the use of the data obtained by the evaluation carried out by three of the human experts that participated in the experiment, since the datasets created from their assessments were highly unbalanced. This imbalance has prevented us from emulating the behavior of different experts and doing comparative studies of great interest, such as creating a decision tree that emulates an average expert evaluator.

However, working with four experts has allowed us to realize that even when working with well-trained professionals, there is a wide disparity of results in the quality assessment. Thus, there is a very high subjectivity factor in this process that hinders the creation of a generalized model. It is important to highlight that this occurs even when working with a single person, since an expert can vary his performance in the evaluation due to subjective factors that vary over time. In particular, the evaluator used for the construction of the decision tree has given only a 89% consistency of results, as we showed in an experiment prior to the study presented here (**Robledano-Arillo**; **Moreno-Pelayo**; **Pereira-Uzal**, 2016).

Since the consistency of results from the best human expert is 89%, one could think that the best effectiveness that we can achieve emulating our human expert will be close to 90%. However, we could obtain a model that emulates the expert's best practice. For example, data could be taken from the intervals of time where his judgments are consistent (that is, he gives the same judgments when repeated images are showed to him). Therefore, the resulting model could perform even better than the expert himself, because the model would not be affected by human limitations such as fatigue, lack of concentration, distractions, and so on.

Finally, we have to emphasize that we are not dealing with a trivial problem of a deterministic nature, but rather with a complex problem of probabilistic type, which would require a more intensive future approximation. That is, it would be necessary to derive a psychometric function that models the quality judgment of the expert. Nevertheless, it has to be pointed out that we have not aspired to create a generalizable psychometric model, but to show that the automated systems that are currently used in the assessment of quality of patrimonial digitizations (based only on the use of color metrics, such a CIE color formulas, that reduce the dimensionality of the perceptual variables of color to a single dimension) do not match well with the perception of quality of human experts. Besides, we want to show that using more complex models, based on n-dimensional vectors and on decision trees, we can obtain better performances, approaching the results of an expert in a particular quality assessment.

5.3. Future research

In future work we plan to try other ML algorithms to give a measure of the quality of the images. Besides, we plan to explore the possibility of emulating the quality judgement given by the best expert from the judgements given by the other experts. Moreover, we want to explore different color representation systems that segment the representation by independent channels of chrominance and luminance such as *Cielab* (*ISO*, 2008) or other perceptual methods.

6. Conclusions

In this paper we have addressed the problem of image quality assessment in the context of cultural heritage digitization by applying machine learning to develop a method that enables the assessment of image quality automatically. More specifically, we have explored the possibility of creating a decision tree that mimics the response of an expert on cultural heritage when observing several images under certain determined and controlled conditions. Different manners of creating the decision tree have been explored. In terms of attributes' selection, we have explored the possibility of creating the classifier using four different selections of attributes: CIE76, CIE76+PSNR+SSIM, CIE76+HSL and CIE76+HSL+-patches. In terms of learning algorithms we have used the well-known C4.5 algorithm by itself, combined with bagging and combined with boosting.

> We have applied machine learning to develop a method that enables the assessment of image quality automatically

The results presented in our work show that decision trees can be used to give a measure of quality in the context of cultural heritage digitization. More particularly, our results show that the best option is to apply boosting to create the classifier that categorizes images as valid or invalid, being the inputs to the learning algorithm CIE76+HSL+patches. That is, being the inputs the average of the 24 color differences between the digitized physical image and real color from the *ColorChecker®* card, the HSL perceptual values (hue, saturation and lightness), and the 24 color differences between the digitized physical image and the real color from each of the 24 patches from the *ColorChecker®* card.

> Decision trees can be used to give a measure of quality in the context of cultural heritage digitization

7. References

Aydin, Tunç-Ozan; **Mantiuk, Rafal**; **Myszkowski, Karol**; **Seidel, Hans-Peter** (2008). "Dynamic range independent image quality assessment". *ACM transactions on graphics,* v. 27, n. 3, article n. 69.
http://resources.mpi-inf.mpg.de/hdr/vis_metric

Brandão, Tomás; **Queluz, Maria-Paula** (2008). "No-reference image quality assessment based on DCT domain statistics". *Signal processing*, v. 88, n. 4, pp. 822-833.
https://doi.org/10.1016/j.sigpro.2007.09.017

Breiman, Leo (1996). "Bagging predictors". *Machine learning*, v. 24, n. 2, pp. 123-140.
https://doi.org/10.1023/A:101805431

Cadík, Martin; **Herzog, Robert**; **Mantiuk, Rafal**; **Myszkowski, Karol**; **Seidel, Hans-Peter** (2012). "New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts". *ACM transactions on graphics*, v. 31, n. 6, article n. 147.
https://doi.org/10.1145/2366145.2366166

Chang, Huiwen; **Yu, Ficher**; **Wang, Jue**; **Ashley, Douglas**; **Finkelstein, Adam** (2016). "Automatic triage for a photo series". *ACM transactions on graphics*, v. 35, n. 4, article n. 148.
https://www.juew.org/publication/sig16-triage.pdf
https://doi.org/10.1145/2897824.2925908

Charrier, Christophe; **Lézoray, Olivier**; **Lebrun, Gilles** (2012). "Machine learning to design full-reference image quality assessment algorithm". *Signal processing: Image communication*, v. 27, n. 3, pp. 209-219.
https://doi.org/10.1016/j.image.2012.01.002

De, Indrajit; **Sil, Jaya** (2011). "No reference image quality assessment using fuzzy relational classifier". In: Deng Hepu; Miao, Duoquian; Lei, Jingsheng; Wang, Fu-Lee (eds.). *Intl conf on artificial intelligence and computational intelligence. AICI 2011. Lecture notes in computer science*, v. 7002, pp. 551-558. ISBN: 978 3 642 23880 2
https://doi.org/10.1007/978-3-642-23881-9_71

Engeldrum, Peter G. (1995). "A framework for image quality models". *Journal of imaging science and technology*, v. 39, n. 4, pp. 312-318.

Engeldrum, Peter G. (2004). "A theory of image quality: The image quality circle". *Journal of imaging science and technology*, v. 48, n. 5, pp. 446-456.
http://www.imcotek.com/pdf_temp/JIST_446-456_04_IQtheory.pdf

Fadgi (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.* Federal Agencies Digitization Initiative; Still Image Working Group.
http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf

Frey, Franzisca S.; **Reilly, James M.** (2006). *Digital imaging for photographic collections: Foundations for technical standards* (2nd ed.). Rochester, NY: Image Permanence Institute.
https://www.imagepermanenceinstitute.org/webfm_send/650

Gao, Xinbo; **Gao, Fei**; **Tao, Dacheng**; **Li, Xuelog** (2013). "Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning". *IEEE Transactions on neural networks and learning systems*, v. 24, n. 12, pp. 2013-2026.
https://doi.org/10.1109/TNNLS.2013.2271356

Gastaldo, Paolo; **Zunino, Rodolfo**; **Heynderickx, Ingrid**; **Vicario, Elena** (2005). "Objective quality assessment of displayed images by using neural networks". *Signal processing: Image communication*, v. 20, n. 7, pp. 643-661.
https://doi.org/10.1016/j.image.2005.03.013

Gastaldo, Paolo; **Zunino, Rodolfo**; **Redi, Judith** (2013). "Supporting visual quality assessment with machine learning". *Eurasip, Journal on image and video processing,* n. 54, pp. 1-15.
https://doi.org/10.1186/1687-5281-2013-54

Hall, Mark; **Frank, Elbe**; **Holmes, Geoffrey**; **Pfahringer, Bernhard**; **Reutemann, Peter**; **Witten, Ian H.** (2009). "The WEKA data mining software: An update". *Sigkdd Explorations*, v. 11, n. 1, pp. 10-18.
https://doi.org/10.1186/1687-5281-2013-54

ISO (2008). *ISO 11664-4:2008* (*CIE S 014-4/E:2007*). *Colorimetry - Part 4: CIE 1976 L*a*b* colour space*. Geneva, Switzerland: International Organization for Standardization.

ISO (2009). *ISO 3664:2009. Graphic technology and photography - Viewing conditions*. Geneva, Switzerland: International Organization for Standardization.

ISO (2012). *ISO 20462-3:2012. Photography - Psychophysical experimental methods for estimating image quality - Part 3: Quality ruler method*. Geneva, Switzerland: International Organization for Standardization.

ISO (2015). *ISO 12646:2015. Graphic technology - Displays for colour proofing – Characteristics*. Geneva, Switzerland: International Organization for Standardization.

ISO (2017a). *ISO 19263-1:2017. Photography - Archiving systems. Part 1: Best practices for digital image capture of cultural heritage material*. Geneva, Switzerland: International Organization for Standardization.
https://www.iso.org/obp/ui/#iso:std:iso:tr:19263:-1:ed-1:v1:en

ISO (2017b). *ISO/TS 19264-1:2017. Photography - Archiving systems -- Image quality analysis - Part 1: Reflective originals*. Geneva, Switzerland: International Organization for Standardization.

Kopf, Johannes; **Kienzle, Wolf**; **Drucker, Steven**; **Kang, Sing-Bing** (2012). "Quality prediction for image completion". *ACM transactions on graphics*, v. 31, n. 6, article n. 131.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.370.1764&rep=rep1&type=pdf
https://doi.org/10.1145/2366145.2366150

Kusuma, Tubagus-Maulana; **Zepernick, Hans-Jürgen** (2003). "A reduced-reference perceptual quality metric for in-service image quality assessment". In: *SympoTIC'03. Joint first Workshop on mobile future and Symposium on trends in communications. IEEE*, pp. 71-74.
https://doi.org/10.1109/TIC.2003.1249092

Larson, Eric C.; **Chandler, Damon M.** (2010). "Most apparent distortion: Full-reference image quality assessment and the role of strategy". *Journal of electronic imaging*, v. 19, n. 1, pp. 1-21.
https://s2.smu.edu/~eclarson/pubs/2010JEI_MAD.pdf
https://doi.org/10.1117/1.3267105

Li, Chaofeng; **Bovik, Alan C.**; **Wu, Xiaojun** (2011). "Blind image quality assessment using a general regression neural network". *IEEE Transactions on neural networks*, v. 22, n. 5, pp. 793-799.
https://live.ece.utexas.edu/publications/2011/Li_tnn_2011.pdf
https://doi.org/10.1109/TNN.2011.2120620

Li, Qiang; **Wang, Zhou** (2009). "Reduced-reference image quality assessment using divisive normalization-based image representation". *IEEE Journal of selected topics in signal processing*, v. 3, n. 2, pp. 202-211.
https://doi.org/10.1109/JSTSP.2009.2014497

Lin, Weisi; **Kuo, C. C. Jay** (2011). "Perceptual visual quality metrics: A survey". *Journal of visual communication and image representation*, v. 22, n. 4, pp. 297-312.
https://doi.org/10.1016/j.jvcir.2011.01.005

Liu, Mohan; **Konya, Iuliu**; **Nandzik, Jan**; **Flores-Herr, Nicolas**; **Eickeler, Stefan**; **Ndjiki-Nya, Patrick** (2012). "A new quality assessment and improvement system for print media". *Eurasip, Journal on advances in signal processing*, v. 2012, n. 109.
https://doi.org/10.1186/1687-6180-2012-109

Liu, Tsung-Jung; **Lin, Weisi**; **Kuo, C. C. Jay** (2013). "Image quality assessment using multi-method fusion". *IEEE transactions on image processing*, v. 22, n. 5, pp. 1793-1807.
https://www.researchgate.net/publication/234047751_Image_Quality_Assessment_Using_Multi-Method_Fusion
https://doi.org/10.1109/TIP.2012.2236343

Luo, Ming-Ronnier; **Cui, Guihua**; **Rigg, Bryan** (2001). "The development of the CIE 2000 colour-difference formula: CIEDE2000". *Color research & application*, v. 26, n. 5, pp. 340-350.
https://www.researchgate.net/publication/229511830_The_development_of_the_CIE_2000_colour-difference_formula_CIEDE2000
https://doi.org/10.1002/col.1049

Ma, Lin; **Li, Songnan**; **Zhang, Fan**; **Ngan, King-Ngi** (2011). “Reduced-reference image quality assessment using reorganized DCT-based image representation”. *IEEE Transactions on multimedia*, v. 13, n. 4, pp. 824-829.
https://doi.org/10.1109/TMM.2011.2109701

Major, John A.; **Mangano, John J.** (1995). “Selecting among rules induced from a hurricane database”. *Journal of intelligent information systems*, v. 4, n. 1, pp. 39-52.
https://doi.org/10.1007/BF00962821

Mantiuk, Rafat; **Kim, Kil-Joong**; **Rempel, Allan G.**; **Heidrich, Wolfgang** (2011). “HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions”. *ACM Transactions on graphics*, v. 30, n. 4, article n. 40.
http://hdrvdp.sourceforge.net/hdrvdp.pdf
https://doi.org/10.1145/2010324.1964935

Mittal, Anish; **Moorthy, Anush-Krishna**; **Bovik, Alan C.** (2012). “No-reference image quality assessment in the spatial domain”. *IEEE Transactions on image processing*, v. 21, n. 12, pp. 4695-4708.
https://doi.org/10.1109/TIP.2012.2214050

Narwaria, Manish; **Lin, Weisi**; **Cetin, A. Enis** (2012). “Scalable image quality assessment with 2D mel-cepstrum and machine learning approach”. *Pattern recognition*, v. 45, n. 1, pp. 299-313.
https://doi.org/10.1016/j.patcog.2011.06.023

Nationaal Archief of the Netherlands (2010). *Digitisation of photographic materials. Guidelines.* September 2010.
https://www.nationaalarchief.nl/sites/default/files/field-file/guidelines_digitisation_photographic_materials.pdf

Oeztireli, A. Cengiz; **Gross, Markus** (2015). “Perceptually based downscaling of images”. *ACM Transactions on graphics*, v. 34, n. 4, article n. 77.
https://people.inf.ethz.ch/~cengizo/Files/Sig15PerceptualDownscaling.pdf
https://doi.org/10.1145/2766891

Ponomarenko, Nikolay; **Jin, Lina**; **Ieremeiev, Oleg**; **Lukin, Vladimir**; **Egiazarian, Karen**; **Astola, Jaakko**; **Vozel, Benoit**; **Chehdi, Kacem**; **Carli, Marco**; **Battisti, Federica**; **Kuo, C. C. Jay** (2015). “Image database TID2013: Peculiarities, results and perspectives”. *Signal processing: Image communication*, v. 30, pp. 57-77.
https://doi.org/10.1016/j.image.2014.10.009

Puglia, Steven; **Reed, Jeffrey**; **Rhodes, Erin** (2004). *Technical guidelines for digitizing archival materials for electronic access: Creation of production master files - Raster images*. U. S. National Archives and Records Administration (NARA).
https://www.archives.gov/files/preservation/technical/guidelines.pdf

Quinlan, J. Ross (1992). *C4.5: Programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN: 978 1 558602380

Robledano-Arillo, Jesús (2016). “25 years of digital conversión, state of the art”. In: *Conservation of photographs: 30 years of science*. Pamplona, Spain: CAAP. ISBN: 978 84 608 4647 5

Robledano-Arillo, Jesús; **Moreno-Pelayo, Valentín**; **Pereira-Uzal, José-Manuel** (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*, v. 39, n. 2.
https://doi.org/10.3989/redc.2016.2.1249
English version on:
https://e-archivo.uc3m.es/handle/10016/23693

Robledano-Arillo, Jesús; **Navarro-Bonilla, Diego** (2017). “Aproximación sistemática a la creación de versiones digitales de negativos fotográficos históricos”. *El profesional de la información*, v. 26, n. 6, pp. 1172-1183.
https://doi.org/10.3145/epi.2017.nov.16

Sheikh, Hamid R.; **Bovik, Alan C.** (2006). “Image information and visual quality”. *IEEE Transactions on image processing*, v. 15, n. 2, pp. 430-444.
https://live.ece.utexas.edu/publications/2004/hrs_ieeetip_2004_imginfo.pdf
https://doi.org/10.1109/TIP.2005.859378

Sheikh, Hamid R.; **Bovik, Alan C.**; **Cormack, Lawrence** (2005). “No-reference quality assessment using natural scene statistics: JPEG2000”. *IEEE Transactions on image processing*, v. 14, n. 11, pp. 1918-1927.
https://www.researchgate.net/publication/3328029_No-reference_quality_assessment_using_natural_scene_statistics_JPEG2000
https://doi.org/10.1109/TIP.2005.854492

Sheikh, Hamid R.; **Sabir, Muhammad-Farooq**; **Bovik, Alan C.** (2006). “A statistical evaluation of recent full reference image quality assessment algorithms”. *IEEE Transactions on image processing*, v. 15, n. 11, pp. 3440-3451.
http://03c91d6.netsolhost.com/PDF/Statistic-of-Full-Reference-UT.pdf

https://doi.org/10.1109/TIP.2006.881959

Suresh, Sundaram; **Babu, R. Venkatesh**; **Kim, Hyoung J.** (2009). "No-reference image quality assessment using modified extreme learning machine classifier". *Applied soft computing*, v. 9, n. 2, pp. 541-552.
https://doi.org/10.1016/j.asoc.2008.07.005

Van-Dormolen, Hans (2012). *Metamorfoze preservation imaging guidelines. Image quality, version 1.0*. National Archives of the Neetherlands. January 2012.
https://www.metamorfoze.nl/sites/metamorfoze.nl/files/publicatie_documenten/Metamorfoze_Preservation_Imaging_Guidelines_1.0.pdf

Wang, Zhou; **Bovik, Alan C.**; **Sheikh, Hamid R.**; **Simoncelli, Eero-Peter** (2004). "Image quality assessment: From error visibility to structural similarity". *IEEE Transactions on image processing*, v. 13, n. 4, pp. 600-612.
http://www.cns.nyu.edu/pub/lcv/wang03-preprint.pdf
https://doi.org/10.1109/TIP.2003.819861

Wang, Zhou; **Simoncelli, Eero P.** (2005). "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model". In: Rogowitz, Bernice E.; Pappas, Thrasyvoulos N.; Daly, Scott J. (eds.). *Proceedings of SPIE. Human vision and electronic imaging X* (18 March 2995), v. 5666, pp. 149-159.
https://doi.org/10.1117/12.597306

Wang, Zhou; **Simoncelli, Eero P.**; **Bovik, Alan C.** (2003). "Multi-scale structural similarity for image quality assessment". In: Matthews, M. B. (ed.). *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, v. 2, pp. 1398-1402.
https://doi.org/10.1109/ACSSC.2003.1292216

Williams, Don; **Longman, Jere** (2003). "Debunking specsmanship: Progress on ISO/TC42 standards for digital capture imaging performance". In: *IS&T PICS Conference*, pp. 77-81.

Zhang, Lin; **Zhang, Lei**; **Mou, Xuanqin**; **Zhang, David** (2011). "FSIM: A feature similarity index for image quality assessment". *IEEE Transactions on image processing*, v. 20, n. 8, pp. 2378-2386.
https://www4.comp.polyu.edu.hk/~cslzhang/IQA/TIP_IQA_FSIM.pdf
https://doi.org/10.1109/TIP.2011.2109730

Zhao, Yang; **Campisi, Patrizio**; **Kundur, Deepa** (2004). "Dual domain watermarking for authentication and compression of cultural heritage image". *IEEE Transactions on image processing*, v. 13, n. 3, pp. 430-448.
https://doi.org/10.1109/TIP.2003.821552