Evaluation of the repeatability and reproducibility levels for colour measurement obtained by digital imaging capture devices

Elísabet Chorro*, Meritxell Vilaseca**, Jorge A. Herrera**, Esther Perales*, Francisco Miguel Martínez-Verdú*, Jaume Pujol**

****Department of Optics, Pharmacology and Anatomy, University of Alicante***
Carretera de San Vicente del Raspeig s/n. 03690 Alicante (Alicante, Spain)

*****Centre for Sensors, Instruments and Systems Development (CD6). Technical University of Catalonia***
Rambla Sant Nebridi 10, 08222 Terrassa. (Barcelona, Spain)

Abstract

Conventional and multi-spectral digital imaging capture devices can be used for colorimetric measurements. However, a preliminary study of their repeatability and reproducibility is very advisable if they want to be used for this purpose. In this work, a methodology to study the repeatability and reproducibility of such systems is proposed and two digital imaging capture devices working as spectroradiometers with three and seven filters, respectively, are analyzed with respect to a reference tele-spectra-colorimeter. The results of the statistical test show that both have good levels of repeatability, although not the same conclusion is reached regarding the reproducibility, due to some systematic errors.

Introduction

Many types of digital imaging capture devices based on digital cameras, such as conventional RGB or multi-spectral cameras, can be used for colorimetric measurements [1-2]. These systems compute the colour coordinates from the digital levels corresponding to the acquisition channels, using either linear or non-linear conversions, or directly from the reconstructed spectra, depending on the performed colorimetric characterization [1, 3].

Most digital imaging capture devices are based on CCD sensors due to its high resolution, high quantum efficiency, wide spectral response, acceptable signal-to-noise ratio, linearity, geometric fidelity, fast response, small size and durability [4-5]. In spite of this, one must bear in mind that they are not perfect detectors, but there are various noise sources inherent to their performance that alter the digital levels corresponding to each pixel, distort the real image acquired in an unknown manner, and diminish the radiometric accuracy, the image quality and the resolution [4].

The specific colorimetric characterization applied as well as the former noise sources may lead to digital imaging capture devices for colorimetric measurements with different repeatability and reproducibility levels. The repeatability [6] is the capability of an instrument for repeating identical measurements under the same conditions and, in general, it can be evaluated as the standard deviation of several measures of the same object. On the other hand, the reproducibility [6-7] is the capability of an instrument for reproducing the expected value when the conditions have changed, for example, when the object, the instrument or the operator are not the same.

Recently, a methodology to evaluate the repeatability and the reproducibility of colour measuring instruments based on the ASTM E2214-08 guidelines [8] has been developed [6-7]. Similar analyses have already been applied to other types of devices, such as multi-gonio-spectrophotometers [9].

The purpose of this work is to extend these methodologies to conventional and multi-spectral digital imaging capture devices used for colour measurements. Specifically, two different systems based on a CCD camera are analyzed: one of them with three colour acquisition channels and another with seven multi-spectral bands. Particularly, this study focuses on evaluating the deviation between the predictions and the chromatic values measured with a reference tele-spectra-colorimeter with exactly the same measurement capture/measurement geometry conditions, in order to avoid possible systematic errors.

Experimental set-up

The configurations of digital imaging capture devices used in this work had been previously developed and characterized [10-11]. Both consisted of a 12 bits cooled monochrome CCD camera (QImaging QICAM Fast1394 12 bit cooled) and a zoom lens (Nikon AF Nikkor 28 – 105 mm). The acquisition channels of each configuration were built with two different sets of filters. Firstly, an RGB liquid crystal tunable filter (Figure 1) was used in the 3-channel colorimetric configuration.

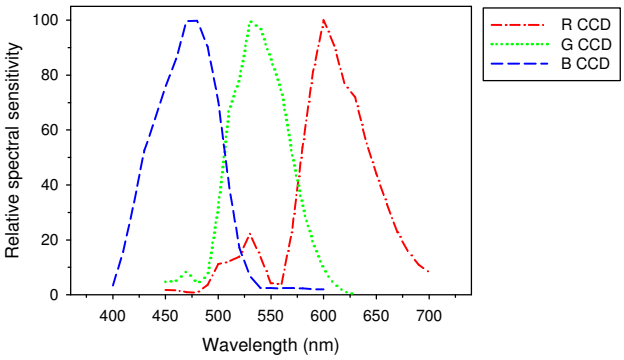

Figure 1. *Relative spectral sensitivities of the RGB channels used in the 3-channel configuration.*

Secondly, a set of seven interference filters with a full width half maximum of approximately 40 nm, covering the whole visible range of the spectrum and fitted in a motorized filter wheel, were used in the 7-channel multi-spectral configuration (Figure 2).

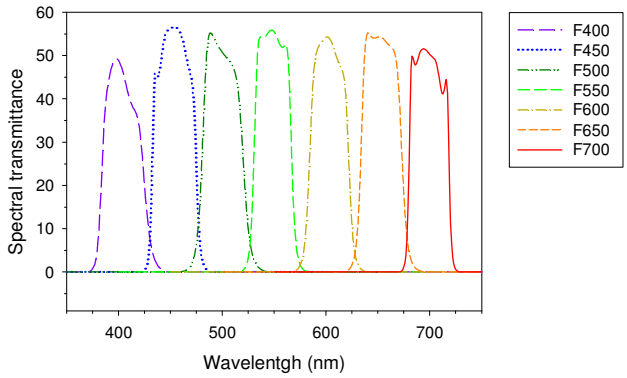

Figure 2. *Transmittance spectra of the seven interference filters used in the 7-channel configuration.*

With the two different configurations, multi-spectral images of the GretagMacbeth ColorChecker DC chart (CCDC, with 180 different colour patches) placed inside a special light booth (63cm x 64cm x 52cm) with a D65 simulator, which provided a diffused and rather uniform illumination over the samples, were acquired (geometry d/0) (Figure 3). Five minutes were used for warming up and stabilizing the illumination system. The acquisitions were repeated 20 times and the averaged digital levels corresponding to a circular area of approximately 1 cm in diameter for each patch were calculated.

Figure 3. *CCDC chart placed inside the light booth used in the experimental set-up and the tele-spectra-colorimeter used as reference instrument.*

Furthermore, the same areas of the colour patches of the Color Checker DC chart were also characterized by means of a tele-spectra-colorimeter (PhotoResearch PR-655 with the MS-75 zoom lens), which provided the spectral radiances ($W/sr \cdot m^2$) in the visible range of the electromagnetic spectrum with exactly the same measurement geometry d/0 used with the digital imaging capture device.

The multispectral system was trained to predict the radiance spectra from the digital levels using a direct transformation that related both sets of values by means of a matrix computed using the Moore-Penrose pseudo-inverse technique [12]. The transformation matrix was calculated taking into account the mean digital levels of each patch from the 20 acquisitions and the corresponding true spectral radiances measured with the tele-spectra-colorimeter.

Finally, the spectral radiances of each colour patch corresponding to the 20 independent acquisitions performed with each configuration of the system (3 and 7 channels) were reconstructed. The CIELAB chromatic values were calculated from the reconstructed radiance spectra, taking into account the CIE-XYZ absolute tristimulus values (in cd/m^2) of all colour samples and the white standard used inside the same scene.

Methods

For an instrument with good degree of repeatability, the chromatic values of all the acquisitions - expressed for example in CIELAB coordinates - would be almost identical, and small deviations among measurements would be caused by random errors. For this reason, the first step for evaluating the repeatability of an imaging capture device is to study if the errors of the chromatic values are random or not. If they are random, it is expected that they can be fitted using a normal distribution. There are a lot of statistical tests for comparing a sample with a reference probability distribution. Particularly, the Kolmogorov-Smirnov-Lilliefors test [13] (K-S-L test) compares a sample with a normal distribution, and it is used to test the null hypothesis with data coming from a normally distributed population, when the null hypothesis does not specify the expected value and variance. If the significance of the test, calculated with 20 degrees of freedom using SPSS or Matlab software with statistical toolbox, is lower than the significance level, usually 0.05 (95 % confidence level), the null hypothesis is rejected, i.e. the error cannot be fitted by a normal distribution, and therefore the statistical deviation cannot be attributed to random errors.

The second step related to repeatability is the study of how many measurements have been averaged for reaching the highest degree of repeatability or, in other words, how the mean and standard deviation are affected by the number of averaged samples. For this issue, one can compute the mean and the deviation using a different number of samples and check when they reach the stability. On the other hand, the hypothesis test is a method for making statistical decisions using experimental data. Specifically, the t-test can be used to compare the mean of a data set made up of less than 20 measurements and its expected value, which is the mean obtained with all data (in our case the 20 measurements). In this case, if the significance of the t-test is lower than the significance level, that is, 0.05, the null hypothesis is rejected; In other words, the means are not equal. Consequently, the average depends on the number of samples averaged. To avoid that, all variables and samples must pass the t-test.

To evaluate the reproducibility of an imaging capture device we analyzed the deviation between the prediction and the measured values with the reference instrument (tele-spectra-colorimeter) in the same viewing position as the digital imaging capture system.

First of all, the partial and total colour differences in the CIELAB colour space among the values measured by the reference instrument and those obtained from the mean of the all measurements performed for each sample were calculated. If the reproducibility level was ideal, all colour differences should be zero. To evaluate if the colour differences are zero or not, the Hotelling test, which is a multivariate test, and the inter-comparison test, which is a univariate test, can be used.

The Hotelling's T^2 metric [7] is an index that measures the tolerance volume (Eq. 1, n = 180) of an instrument for a given statistical significance in terms of the partial colour differences, ΔL^*, Δa^* and Δb^*, and the null hypothesis is that the colour differences between the same samples measured by the reference and the test instruments are zero. Generally, one rejects the null hypothesis if the p-value is smaller than the significance level, that is, 0.05 (95% confidence level). These multivariate statistical method has already been used by other authors with the same purpose [7, 9] and it has also been implemented in Matlab [14].

$$S = \begin{bmatrix} \mathrm{var}(\Delta L^*) & \mathrm{cov}(\Delta L^*, \Delta a^*) & \mathrm{cov}(\Delta L^*, \Delta b^*) \\ \mathrm{cov}(\Delta L^*, \Delta a^*) & \mathrm{var}(\Delta a^*) & \mathrm{cov}(\Delta a^*, \Delta b^*) \\ \mathrm{cov}(\Delta L^*, \Delta b^*) & \mathrm{cov}(\Delta a^*, \Delta b^*) & \mathrm{var}(\Delta b^*) \end{bmatrix} \quad (1)$$

$$T^2 = n \cdot \left[\Delta L^* \quad \Delta a^* \quad \Delta b^*\right]^T \cdot S^{-1} \cdot \left[\Delta L^* \quad \Delta a^* \quad \Delta b^*\right] \quad (2)$$

Additionally, the inter-comparison test [7], which is an univariate test derived from propagation of errors and the Chi-squared statistical distribution, was calculated (Eqs. 3-5). In this case, if the total colour difference average of the differences (ΔE) is higher than the critical value ($t_{\Delta E}$), which is the statistical parameter of this test, the difference is statistically significant.

$$\alpha = \frac{\mathrm{mean}\left(\Delta L^*\right)}{\mathrm{mean}\left(\Delta E_{ab}\right)}, \quad \beta = \frac{\mathrm{mean}\left(\Delta a^*\right)}{\mathrm{mean}\left(\Delta E_{ab}\right)}, \quad \gamma = \frac{\mathrm{mean}\left(\Delta b^*\right)}{\mathrm{mean}\left(\Delta E_{ab}\right)} \quad (3)$$

$$g_E = g_{11}\alpha^2 + g_{22}\beta^2 + g_{33}\gamma^2 + 2g_{12}\alpha\beta + 2g_{23}\beta\gamma + 2g_{13}\alpha\gamma \quad (4)$$

$$t_{\Delta E} = \sqrt{\frac{\chi_3^2}{n \cdot g_E}}, \quad n = 180 \quad (5)$$

In both tests, statistically significant differences would imply that differences among the values measured by the reference instrument and those obtained from measurements of the digital imaging capture device are due to systematic errors, but not exclusively to random errors. However it is impossible to make a distinction between types of the systematic errors only with the statistical test of reproducibility. Other tests [15-16] are needed in order to achieve that, although it is not the aim of this work.

Finally, the study of the reproducibly was repeated, but performing a comparison with colour differences between the chromatic values obtained by the two configurations of the digital imaging capture devices. In this case, the purpose of the test was to evaluate if there were differences between the results obtained by both devices, in order to establish which one would be preferred for colour measurement. If no differences exist, the device with a better repeatability or with a higher ease of use, would be a better choice.

Results

The 20 CIELAB chromatic values, calculated for each patch by the 3-channel and 7-channel configurations were used as variables in the Kolmogorov-Smirnov-Lilliefors test with 20 degrees of freedom. The results corresponding to four representative patches are shown in Table I.

Table I: Significance of the Kolmogorov-Smirnov-Lilliefors test for 4 of the 180 patches using the two configurations of the digital image capture device.

		B2	C3	D4	C6
3-channel	**L***	0.122[a]	0.200[a]	0.053[a]	0.020
	a*	0.200[a]	0.200[a]	0.200[a]	0.200[a]
	b*	0.200[a]	0.200[a]	0.200[a]	0.730[a]
7- channel	**L***	0.200[a]	0.200[a]	0.200[a]	0.690[a]
	a*	0.200[a]	0.200[a]	0.200[a]	0.200[a]
	b*	0.100[a]	0.200[a]	0.200[a]	0.140[a]

[a] The limit has been reached and so these values are not tabulated (they are higher)

If the significance of the test, shown in table I, is lower than 0.05, the null hypothesis is rejected. For instance, the value L* using the 3-channel configuration for the patch C6 cannot be fitted with a normal distribution. Table II summarizes how many samples did not pass the normality test for the three variables.

Table II: Summary of Kolmogorov-Smirnov-Lilliefors test for the 180 patches using the two configurations of the digital image capture device.

		Nº of samples No Pass	% No Pass
3-channel	**L***	9	5.3%
	a*	8	4.7%
	b*	8	4.7%
7- channel	**L***	16	9.4%
	a*	10	5.9%
	b*	11	6.5%

It can be seen that more than 90% of the patches have chromatic values which can be fitted with a normal distribution and therefore, it is expected that their deviations are caused by random errors. However, taking into account the L* value, 3 samples did not pass the test with any instrument, The same analysis performed with the a* value reveals that all samples passed meanwhile using the b* value, 2 samples did not pass.

In order to study how the mean and standard deviation are affected by the number of averaged samples, these parameters

are plotted versus the number of samples averaged in figures 4, 5 and 6 for the patch B2.

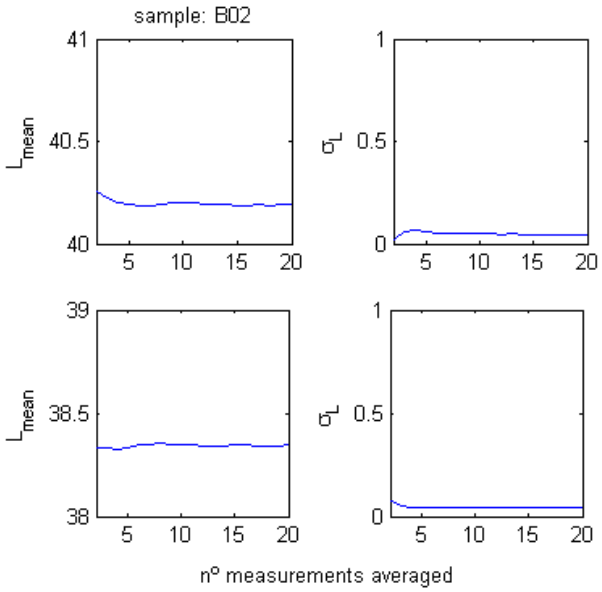

Figure 4. *Mean and deviation versus number of averaged samples of the L* for the B2 patch measured using the 3-channel configuration (top) and 7-channel configuration (bottom) of the digital imaging capture device.*

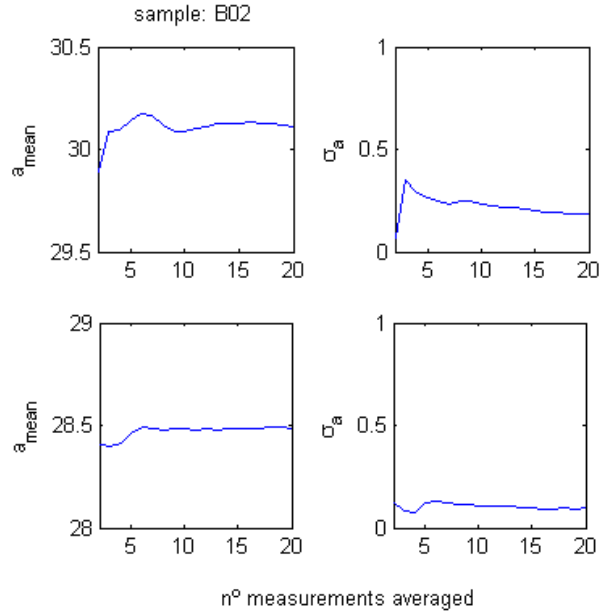

Figure 5. *Mean and deviation versus number of averaged samples of the a* for the B2 patch measured using the 3-channel configuration (top) and 7-channel configuration (bottom) of the digital imaging capture device.*

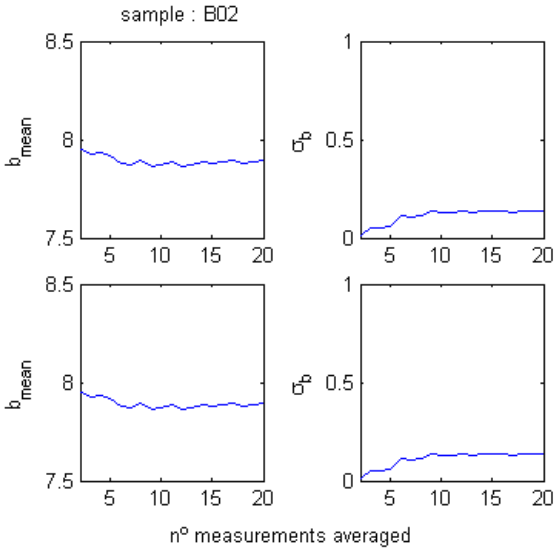

Figure 6. *Mean and deviation versus number of averaged samples of the b* for the B2 patch measured using the 3-channel configuration (top) and 7-channel configuration (bottom) of the digital imaging capture device.*

The obtained results suggest that stability is reached from 7 to 10 averaged measurements. To reinforce this conclusion, the statistical hypothesis test is used in order to decide how many samples are needed to compute the average. The null hypothesis is that the mean of the distribution with less than 20 measurements has the same mean that the expected value, in our case, the mean obtained with 20 measurements. The number of samples that do not pass the t-test using this null hypothesis with a significance level of 0.05 is the indicator to decide the minimum number of samples for averaging. The results obtained in this case are shown in bar plots of the figures 7, 8 and 9.

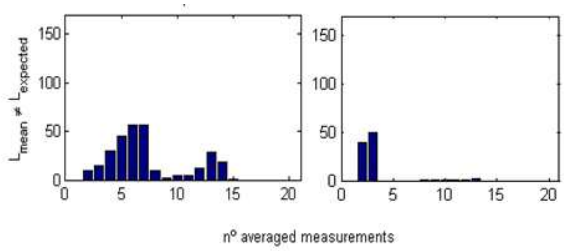

Figure 7. *Bar plot showing the number of samples that do not pass the t-test for L* value measured using the 3-channel configuration (left) and the 7-channel configuration (right) of the digital imaging capture device.*

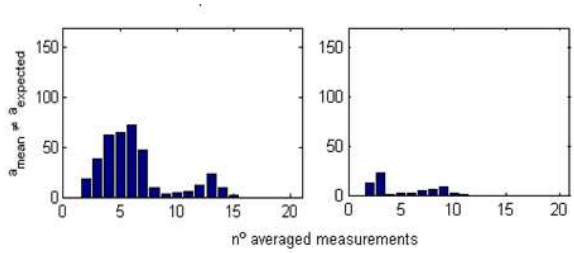

Figure 8. *Bar plot showing the number of samples that do not pass the t-test for a* value measured using the 3-channel configuration (left) and the 7-channel configuration (right) of the digital imaging capture device.*

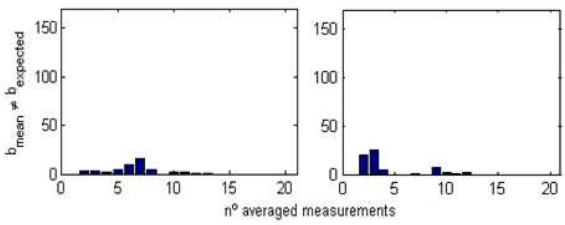

Figure 9. *Bar plot showing the number of samples that do not pass the t-test for b* value measured using the 3-channel configuration (left) and the 7-channel configuration (right) of the digital imaging capture device.*

Figures 7,8 and 9, specifically bar plots on the left, show a relative minimum around 9 averaged samples meanwhile bar plots on the right show it around 4 or 5. However, the best results are obtained with an average of 16 measurements in both configurations of the digital imaging capture device, since from this point on there is not any colour patch whose average is not equal to the total average or expected value. It can be concluded that at least 10 measurements are necessary to compute the average obtaining quite reasonably good results related to repeatability.

Regarding the reproducibility obtained by the digital imaging capture system analyzed, Table III summarizes the results when the Hottelling test is applied to both configurations. In this case, the p-values of both configurations are lower and therefore, the null hypothesis is rejected.

Table III: Hottelling test between the two configurations of the digital imaging capture device and the reference instrument.

	3-channel	7-channel
Sample size	180	180
Degrees of freedom	3	3
T^2	107.4397	72.2556
p-value	0.0000	0.0000

The results of the inter-comparison test are shown in table IV. If the average of the colour difference ΔE_{ab} is higher than the critical value $t_{\Delta E}$, the difference is statistically significant.

Equivalently to the Hotelling test, the results show statistically significant differences among the values measured by the reference instrument and those obtained from measurements of the 3-channel and 7-channel configurations of the digital imaging capture device, since the colour differences ΔE_{ab} are higher than the critical value $t_{\Delta E}$. Therefore, the differences found are related to systematic errors, and not exclusively to random errors.

Table IV: Inter-comparison test between the two configurations of the digital imaging capture device and the reference instrument.

	3-channel	7-channel
$t_{\Delta E}$	1.92	1.33
ΔE_{ab}	5.20	2.96

Finally, the study of the reproducibly between inter-instruments is summarized in the Tables V and VI, that is, using the Hottelling and the inter-comparison tests, respectively. However, in this case the comparison is performed between the two configurations of the digital imaging capture devices.

Table V: Hottelling test between the two configurations of the digital imaging capture devices.

Sample size	180
Degrees of freedom	3
T^2	18.5508
p-value	0.0003

Table VI: Inter-comparison test between the two configurations of the digital imaging capture device.

$t_{\Delta E}$	3.82
ΔE_{ab}	4.31

In both tests, the results show statistically significant differences among the values measured by the 3-channel and 7-channel configurations of the digital imaging capture device.

Conclusions

A methodology for evaluating the repeatability and the reproducibility for colour measurement obtained by digital imaging capture devices has been presented in this work. Specifically, two configurations with three and seven acquisition channels of a CCD-based digital imaging system have been evaluated, The study of the repeatability was made adapting traditional statistical tests; Kolmogorov-Smirnov-Lilliefors test was used to study the normality of the measurements, and the t-test to set the minimal number of measurements needed for the average. The study of the reproducibility was performed accordingly to the ASTM rules and other authors recommendations, such as the Hotelling test, which is a multivariate test, and the inter-comparison test, which is an univariate test, in order to evaluate if the colour differences were zero or not.

Regarding the results obtained by the Kolmogorov-Smirnov-Lilliefors test, it can be concluded that the majority of CIELAB chromatic values of the 180 patches can be fitted using a normal distribution, once applied the same spectral and colorimetric characterization for both configurations of the digital imaging capture device.

Concerning the minimum number of samples, it can be seen that at least 10 measurements are necessary to compute the average and, therefore, obtaining reasonably good results related to repeatability. However, the t-test results suggest using 16 measurements in both configurations.

The results of reproducibility show that there are statistically significant differences among the values measured by the reference instrument and those obtained with the two configurations of the digital imaging capture system, even though the measurement/capture geometries used as well as the light source were exactly the same. Therefore, it is shown that the differences found are caused by systematic errors, and not exclusively by random errors.

Moreover, the study of the reproducibly among the two configurations of the digital imaging capture device reaches the same conclusion, i. e. there are significant differences between the results obtained by them .

Future work is oriented to extend the use of the former statistical tests to evaluate the repeatability and reproducibility levels for spectral measurements rather than colorimetric.

References

[1] Sharma, G. *Digital color imaging handbook*. CRC Press, Boca Raton, FL, 2003.

[2] Westland, S. and Ripamonti, C. *Computational colour science using MATLAB*. J. Wiley, Hoboken, NJ, 2004.

[3] Tzeng, D.-Y. and Berns, R. S. A review of principal component analysis and its applications to color technology. *Color Research & Application*, 30, 2 (2005), 84-98.

[4] Janesick, J. R. *Scientific charge-coupled devices*. SPIE Press, Bellingham, Wash., 2001.

[5] Holst, G. C. *CCD Arrays, Cameras and Displays*, Bellingham 1998.

[6] Wyble, D. R. and Rich, D. C. Evaluation of methods for verifying the performance of color-measuring instruments. Part I: Repeatability. *Color Research & Application*, 32, 3 (2007), 166-175.

[7] Wyble, D. R. and Rich, D. C. Evaluation of methods for verifying the performance of color-measuring instruments. Part II: Inter-instrument reproducibility. *Color Research & Application*, 32, 3 (2007), 176-194.

[8] ASTM *Standard Practice for Specifying and Verifying the Performance of Color-Measuring Instruments*. West Conshohocken, 2008.

[9] Chorro, E., Perales, E., Navarro, V., Alcón, N., Rabal, A. and Martínez-Verdú, F. M. *Reproducibility comparison between multi-gonio-spectophotometers*. Colour Society of Australia, Inc., Sydney, 2009.

[10] de Lasarte, M.., Pujol, J., Arjona, M. and M., V. *Influence of the Size of the Training Set on Colour Measurements Performed Using a Multispectral Imaging System*. Proc. IS&T Fourth European Conference on Colour in Graphics, Imaging and Vision (Terrassa, Spain), pg. 437 (2008).

[11] Vilaseca, M., Mercadal, R., Pujol, J., Arjona, M., de Lasarte, M., Huertas, R., Melgosa, M. and Imai, F. H. Characterization of the human iris spectral reflectance with a multispectral imaging system. *Appl Opt*, 47, 30 (2008), 5622-5630.

[12] Vilaseca, M., Pujol, J., Arjona, M. and de Lasarte, M. Multispectral system for reflectance reconstruction in the near-infrared region. *Appl Opt*, 45, 18 (2006), 4241-4253.

[13] Lilliefors, H. W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *J. Am. Stat. Assoc.*, 62 (1967), 399-402.

[14] Trujillo-Ortiz, A. *HotellingT2,* http://www.mathworks.com/matlabcentral/fileexchange/2844.

[15] Berns, R. S. and Reniff, L. An abridged technique to diagnose spectrophotometric errors. *Color Research & Application*, 22, 1 (1997), 51-60.

[16] Campos Acosta, J., Rubiño López, M., Castillo Rubí, F. J. and Pons, A. Colour Measurement Instruments Comparison. *Óptica Pura y Aplicada*, 37, 1 (2004), 113-118.

Author Biography

Elísabet Chorro received her BS in Physics (Optics branch) from the University of Valencia at Valencia in 2003 and her MSc Dissertation on Physics from the Department of Physics, Systems Engineering and Signal Theory at the University of Alicante (Alicante, Spain) in 2006. Since 2004 she works with the Color and Vision Group of the University of Alicante. Her work has primarily focused on Industrial Colorimetry, Color Vision and Color Imaging.

Acknowledgements

This research was supported by the Spanish Ministry of Science and Innovation under grants DPI2008-06455-C02-01 and DPI2008-06455-C02-02.