Towards a psychophysical evaluation of colour constancy algorithms

Javier Vazquez, Maria Vanrell, Ramon Baldrich and C.Alejandro Párraga; Centre de Visió per Computador, Computer Science Department, Universitat Autònoma de Barcelona, Edifíci O, Campus UAB (Bellaterra), C.P.08193, Barcelona, Spain

Abstract

Computational colour constancy tries to solve the problem of recovering the illuminant of a scene from an acquired image. The most popular algorithms developed to deal with this problem use heuristics to select a unique solution from within the feasible set. Their performance has shown that there is still a long way to go to globally solve this problem as a preliminary step in computer vision. Recent works tried to insert high-level constraints to improve the selection step, whose plausibility could be evaluated according to their performance on the final visual task. To allow comparisons of constraints independently of the task, in this work we present a new performance measure, the perceptual angular error. It tries to evaluate the performance of a colour constancy algorithm according to the perceptual preferences of humans instead of the actual optimal solution. To this end, we present a new version of our "MaxName" algorithm, which aims at solving the illuminant problem using high-level information such as the number of identifiably colours on a scene. Afterwards, we show the results of a psychophysical experiment comparing three colour constancy algorithms. Our results show that in more than half of the judgements the preferred solution is not the one closest to the optimal solution. This makes us conclude that such a perceptual comparison is feasible, and we could benefit from the construction of a large colour constancy database of calibrated images, labelled according to the illuminant preferred by human observers.

Introduction

Colour Constancy is the ability of the human visual system to perceive a stable representation of colour despite illumination changes. Like other perceptual constancy capabilities of the visual system, colour constancy is crucial to succeed in many ecologically relevant visual tasks such as food collection, detection of predators, etc. The importance of colour constancy in biological vision is mirrored in computer vision applications, where success in a wide range of visual tasks relies on achieving a high degree illuminant invariance degree. In the last twenty years, research in computational colour constancy has tried to solve the problem of recovering the illuminant of a scene from an acquired image. Although this is a problem effortless solved by the visual system, it has been shown to be a mathematically ill-posed problem which therefore does not have a unique solution.

A common computational approach to illuminant recovery (and colour constancy in general) is to produce a list of possible illuminants (feasible solutions) and then use some assumptions, based on the interactions of scene surfaces and illuminants to select the most appropriate solution among all possible illuminants. A recent extended review of computational colour constancy methods was provided by Hordley in [1]. In this review, computational algorithms were classified in five different groups according to how they approach the problem. These were (a) simple statistical methods [2], (b) neural networks [3], (c) gamut mapping [4,5], (d) probabilistic methods [6] and (e) physics-based methods [7]. Comparison studies ([8], [9]) have ranked the performance of these algorithms, which usually depend on the properties of the image dataset and the statistical measures used for the evaluation. It is generally agreed that, although some algorithms may perform well in average, they may also perform poorly for specific images. This is the reason why some authors [10] have proposed a one-to-one evaluation of the algorithms on individual images. In this way, comparisons become more independent of the chosen image dataset. However, the general conclusion is that more research should be directed towards a combination of different methods, since the performance of a method usually depends on the type of scene it deals with [11]. Recently, some interesting studies have pointed out towards this direction [12], i.e. trying to find which statistical properties of the scenes determine the best colour constancy method to use. In all these previous approaches, the evaluation of the performance of the algorithms has been based on computing the *angular error* between the selected solution and the actual solution that is provided by the acquisition method.

Other recent proposals [13, 14] turn away from the usual approach and deal instead with multiple solutions "delegating" the selection of a unique solution to a subsequent step that depends on high-level, task-related interpretations, such as the ability to annotate the image content. In this example, the best solution would be the one giving the best semantic annotation of the image content. It is in this kind of approaches where the need for a different evaluation emerges, since the performance depends on the visual task and this can lead to an inability to compare different methods. Hence, to be able to evaluate this performance and to compare it with other high-level methods, in this paper we propose to explore a new evaluation procedure.

Thus, the goal of this paper is twofold, firstly we address the problem of evaluating colour constancy methods using psychophysical data instead of the usual angular error from the optimal solution, and secondly we present a simpler version of the algorithm of Tous [13], MaxName, and its evaluation with this new approach. In the last section we discuss the results and we outline how a global dataset should be built in order to be able to achieve this perceptual evaluation of colour constancy algorithms.

Perceptual performance evaluation

Assuming the ill-posed nature of the problem and the difficulty of finding the optimal solution, we propose to bring the computational colour constancy algorithms towards a simulation of human colour constancy abilities by trying to match *computational* solutions to *perceived* solutions. Hence, in this paper we propose a new evaluation measurement, the

Perceptual Angular Error, which is based on perceptual judgements of adequacy of a solution instead of the physical solution. This work gives a preliminary approach towards what we call a perceptual evaluation of computational colour constancy algorithms.

The approach that we propose in this work does not try to give an alternative line research to the current trends, which focus on classifying scene contents to efficiently combine different methods. Here we try to complement these efforts from a different point of view we could consider as more on a top-down direction, instead of the bottom-up nature of the usual research.

Differences between colour constancy algorithms essentially rely on two different aspects: (a) the assumptions made on the scene properties (such as grey-mean content of the scene, existence of a white patch, or highlights, etc.) or (b) the constraints on the recovered image (maximum global intensity as in MV C-Rule, maximum number of identifiable colour names, etc.). In other cases, assumptions and constraints are combined providing interesting approaches based on the use of most likely surfaces and illuminants (as in color by correlation or Bayesian colour constancy). From this point of view, in this work we point out that performance evaluation of different algorithms is intimately related to specific considerations of the method nature. If we want to measure the relevance of an assumption made on the scenes, we will need to evaluate on what kind of scenes the algorithm performs better, meanwhile if we are trying to evaluate the plausibility of a constraint in the selection of the best solution, then other perceptual measures could be more suited. In this work we focus on this second approach, and we propose to evaluate the adequacy of top-down constraints on CC methods by evaluating their correlation with the human colour constancy preferences, instead of their agreement with the physical solutions. We will show in the results section that human preferred solutions do not clearly match with the optimal solutions.

As mentioned before, the most common performance evaluation for colour constancy algorithms consists in measuring how close their proposed solution is to the optimal solution, independently of the goal they are trying to deal with. This has been computed as

$$e_{ang} = a\cos\left(\frac{\rho_w \hat{\rho}_w}{\|\rho_w\|\|\hat{\rho}_w\|}\right) \quad (1)$$

which represents the angle between the actual white point of the scene illuminant, ρ_w, and the estimation of this point given by the colour constancy method, $\hat{\rho}_w$, which can be understood as a chromaticity distance between the physical solution and the estimate. The current consensus is that none of the current algorithms present a good performance on all the images [15], and a combination of different algorithms offers a promising option for further research. Our proposal, here is to introduce a new measure, the *perceptual angular error*, $e^p{}_{ang}$, that would be computed in a similar way, using ρ_w^p as the perceived white point of the scene measured psychophysically from the image observation and an estimation of this point given by the colour constancy method, $\hat{\rho}_w^p$.

Now, the problem is to define what we mean by "preferred illuminant" of a scene. Below we present a preliminary study towards this objective and finally outline how to build a wide-ranging image dataset to evaluate computational colour constancy within this framework

To sum up, we can state that studying the statistical properties of images to improve the algorithms from a bottom-up perspective (i.e. to improve the performance by approximating the scene information to the algorithm assumptions) has been the prevalent methodology in the literature so far. Here, we propose an approach to the problem from the opposite (top-down) direction using a methodology which validates the performance of the selection algorithm, that is the plausibility of a selection constraint, by correlating its behaviour with what human colour constancy does. This approach starts by proposing a general methodology to evaluate the suitability of high-level constraints by comparing them, independently of the visual task for which they were defined. Otherwise, we can have a wide range of different evaluations of plausible constraints based on the task for which they were designed, such as, their efficiency for image annotation (as in [14]), object recognition or tracking, as it has been usually done to validate illuminant-invariant normalisations [16].

In this paper, we give a preliminary step towards this goal by computing an estimate of this perceptual angular error on three different types of algorithms. We have selected a simple algorithm based on a scene assumption, that is the Grey-world method [2], another algorithm based on a appearance constraint as it is the Maximum Volume C-Rule [17], and another method based on a high-level constraint, the Nameability, that is introduced in the next section.

Maximum Nameability: a high-level constrained method

A recent proposal [13] suggested that it was possible to find an answer to the colour constancy problem by considering multiple solutions generated by the so called *nameability assumption*. This assumption is based on the idea that weighting the solutions accordingly to their ability to assign colour names to the image content would make sense in a global visual task of image annotation. The colour name assignment was done by the computational model of Benavente-*et al* [18], where a fuzzy system allows assigning quantitative fuzzy names to any colour point in an image.

In this paper we propose a simple version of this approach which selects a unique solution from the weighted feasible set, that is the solution that assigns known colour names to a maximum number of points in the image. We will refer to it as the MaxName algorithm. The selection of this unique solution will allow comparing the efficiency of this nameability assumption with other known colour constancy algorithms, since one of the goals of this paper is to set up a new framework to evaluate the plausibility of colour constancy constraints from a psychophysical point of view.

Here, we briefly summarise the main steps of this MaxName algorithm, which is based on building the prior information of the "nameable" colours which are given by

$$\mu_k = \int_\omega S(\lambda)E(\lambda)R_k(\lambda)\partial\lambda \quad , k=R, G, B \quad (2)$$

where, $S(\lambda)$ are the suface reflectances having maximum probability of being labeled with a basic colour name (from the work of Benavente, which are called focal reflectances), in addition we added a set of skin reflectances. $E(\lambda)$ is the power distribution of a D65 illuminant and $R_k(\lambda)$ are the CIE RGB 1955 Colour Matching Functions.

We define μ as the set of all k-dimensional nameable colours obtained from equation (2). The number of elements of μ will depend on the number of reflectances used. We will compute the *Semantic Matrix*, denoted as *SM*, which is a binary representation of the colour space as a matrix where a point in is set to 1 if it represents a "nameable" colour, that is belonging to μ, and 0 otherwise.

Then, for a given input image, *I,* we will compute all possible illuminant changes $I_{\alpha,\beta,\gamma}$. For each $I_{\alpha,\beta,\gamma}$ its nameability value is computed by counting how many points of the mapped image are "nameable" colours in *SM*. Computationally, this process can be done by a correlation in a log space:

$$Nval_{\alpha,\beta,\gamma} = \log(H_{bin}(I)) * \log(SM) \quad (3)$$

where H_{bin} is the binarized histogram of an image. *Nval* at the position (α,β,γ) is the number of coincidences between the SM and $I_{\alpha,\beta,\gamma}$.

Nval is a 3-dimensional matrix, depending on all the feasible maps, (α,β,γ). From this matrix, we will select the most feasible illuminant as the one that accomplishes:

$$(\alpha,\beta,\gamma) = \arg\max_{(\alpha,\beta,\gamma)} Nval \quad (4)$$

that is, the one giving maximum number of nameable colours.

Experiment

Psychophysical experiments were performed using two different image databases. Database A consists of 21 scenes, each one acquired under four different illuminants, taken from the Simon Fraser database [19], totalling 84 test images. Database B consisted of 60 scenes acquired using a calibrated Sigma Foveon D10 digital camera under natural illumination (recorded between 2:00pm and 5:00pm around Barcelona city). The camera colour sensors' spectral sensitivities were measured using a set of 31 spectrally narrowband interference filters and a TopCon SR1 telespectroradiometer (in a process similar to that used in [20,21]). All the pictures in database B included a 18% reflectance grey card, which allows us to manipulate the colour of the illuminant. The pictures were digitally "reilluminated" using 5 different illuminants (three standardised illuminants: 4000K, 7000K, 10000K, and two arbitrary illuminants: *Yellowish* and *Purplish*), totalling 300 test images.

We applied the three colour constancy algorithms (the Gray-World, the C-Rule with the maximum volume selection criteria and our proposed MaxName method based on the maximum nameability constraint) on both image databases getting one solution per test image per algorithm. These solution (i.e. illuminant-removed) images were converted from their original colour space (CIERGB or Device-dependent RGB space) to CIEXYZ and were presented on a calibrated CRT monitor (Viewsonic P227f) using a digital video processor (Cambridge Research Systems Bits++). Experiments were conducted in a dark room.

Experiment 1 was conducted using database A. For this experiment there were 10 naïve observers recruited among Barcelona university students and staff (none of the observers had previously seen the picture database). Pairs of pictures (each obtained using one of the two colour constancy algorithms) were presented side by side on a grey background (31 Cd/m^2). Each picture was viewed from 146 cm and subtended 7.5 x 5.1 deg to the observers.

Experiment 2 was conducted using database B and similar viewing conditions except that the pictures were presented one on top of the other instead of side by side and subtended 10.5 x 5.5 degrees to the observer.

After each presentation, observers were asked to select the picture that seemed most "natural". There was no time limit, although observers were encouraged to spend about 20 second per picture. The presentation order was randomised.

The term "natural" was chosen not because it refers to "natural objects" but because it refers to natural viewing conditions, implying the least amount of digital manipulation. Figure 1 shows some exemplary pictures from database B. The pictures on the left are examples of images chosen as "natural" most of the time and while those on the right are examples of images hardly ever selected as "natural".

***Figure 1**: Images selected as natural (left) and no selected (right)*

Results and discussion

Table 1 shows a summary of the results of Experiment 1 (using database A). Each row shows the percentage of choices of each method when compared to the others (columns). For example, C-Rule was chosen 53.1% of the time against Grey-World and 50.2% of the time against MaxName.

Table 1: Comparison among different methods (Database A)

	C-Rule	Grey-World	MaxName
C-Rule	-	53.1%	50.2%
Grey-World	46.9%	-	45.8%
MaxName	49.8%	54.2%	-

One of the main criticisms to the use of Database A, is that the objects and scenes depicted and the illumination are highly artificial (i.e. not representative of the real world). This was the main reason why we repeated all our experiments using a second database (Database B) consisting mostly of naturalistic scenery under natural illumination with real everyday objects. The analysis that follows is based on these images.

Results from both experiments show that subjects considered the image closer to the optimal (physical) solution to be the most "natural" only about 50% of the time. This shows that the smallest angular error relative to the optimal physical solution (which is the preferred measure for many such comparisons) does not agree with the solution favoured by our observers. **Figure 2** shows some examples of these selections.

Figure 2: *Psychophysical choices (left) and their minimum angular error images (right)*

As mentioned before, it is the current consensus that none of the algorithms present a good performance on all the images. Our results reflect also this fact. Hence, the performance seems to be dependent on the image content. **Figure 3** shows that perceptual angular error outlines some kind of clustering dependent on the scene contents. The results of the experiment show that some colour constancy methods performs better in some scenes and underperform in others. For example, in sky-forest scenes C-Rule performs better, and, as it is logical, in images that have mean grey Grey-World performs the best. However, this conclusion must be confirmed with a high number of scenes.

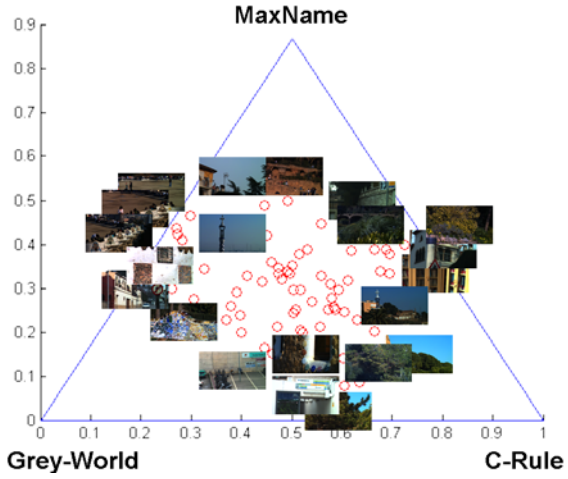

Figure 3: *Some scenes classified by their best method.*

Since our experiment was based on a series of one-to-one image comparisons, we lack the absolute reference data necessary to calculate the exact perceptual angular error. However, it is possible to give an estimation from the one-to-one comparisons obtained from our experiments. This approximate measure was obtained by computing the angular distance between each solution and the psychophysically selected solution for each test image. **Figure 4** shows this estimation of the perceptual angular error on the y-axis, versus an ordered ranking of the images in Database. The area under the curves in **Figure 4** represents a general measure of this perceptual angular error. These areas are: CRule = 7,2442º, MaxName = 6,5620º and Grey-World = 11,9071º

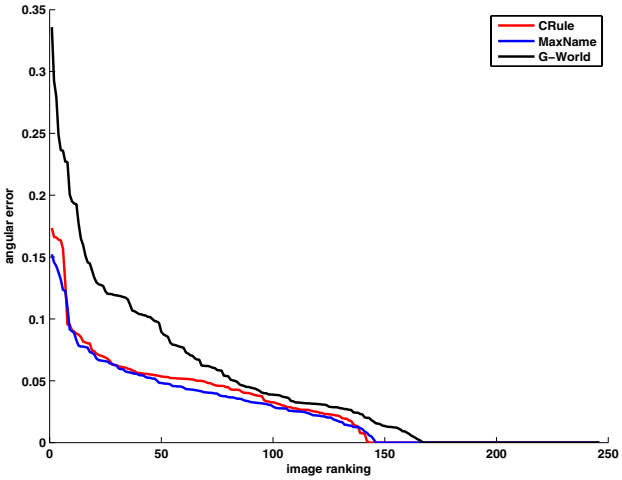

Figure 4: *Perceptual angular error estimation based on the one-to-one experiment.*

To be able to compute an absolute measure of the perceptual angular error, it would be necessary to build an image database where each picture would be linked to the preferred illuminant selected by observers (similarly to what has been done in [22] for paintings). .

Figure 5 shows a histogram of the number of pictures as a function of perceptual angular error for each method. This is a more general plot to compare the algorithms accordingly with this new measure. It shows that CRule and MaxName have similar behaviours while there are significantly less images with low perceptual angular error produced by the Grey-World method.

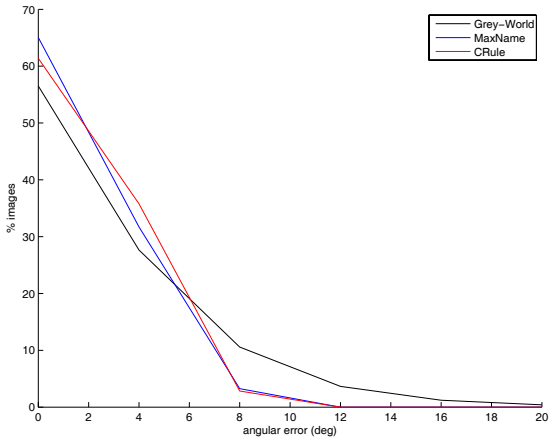

Figure 5: *Percentage of images selected by each algorithm with its corresponding perceptual angular error.*

Table 2: Experimental results for database B

Method	Wins
C-Rule	26.67%
Grey-World	19.0%
MaxName	23.33%
C-Rule-MaxName	5,33%
C-Rule-GreyWorld	2,67%
MaxName-GreyWorld	5%
3-equally selected	18%

Table 3: Comparison among different methods (Database B)

	Wins	Draws with another method	Total
C-Rule	26.67%	8.0%	34.67%
Grey-World	19.0%	7.67%	26.67%
MaxName	23.33%	10.33%	33.66%

Table 2 and Table 3 summarize how each method acts against the others in terms the number of times it has been selected. In Table 2 we list all the cases separately, here we have to highlight that there are some cases (18%) where all three methods have been equally selected by observers and these are not represented in Table 3. In this last table, we show how each method behaves. For each algorithm we consider its performance individually or in pairs. We can see that the most highly selected methods are the C-Rule and MaxName while Grey-World was selected significantly less.

Conclusion

This work is a preliminary step towards a psychophysical evaluation of colour constancy algorithms; which also aims at exploring the high-level constraints needed for the selection of a feasible solution. We have shown the results of a psychophysical experiment that gives a first estimate of the perceptual angular error, trying to measure the proximity of the computational solutions versus the solutions of the human colour constancy.

Within this framework we have evaluated three computational colour constancy algorithms: C-Rule, Grey-world and MaxName. This last one has been briefly presented in section 4 as a simplified version of the Tous algorithm [13]. MaxName algorithm is based on a high-level constraint that estimates the illuminant accordingly with the ability of giving basic colour names to the image content.

The results of the experiments show that in half of the judgments, subjects have preferred solutions that are not the closest ones to the optimal solutions. C-Rule and MaxName methods have shown similar results which are better than the selections given by the Grey-world method. But, the main conclusion is that further work should be done in the line of building a large dataset of images linked to the perceptually preferred judgments.

Acknowledgements

This work has been partially supported by projects TIN2004-02970, TIN2007-64577 and Consolider-Ingenio 2010 CSD2007-00018 of Spanish MEC (Ministery of Science). CAP was funded by the Ramon y Cajal research programme of the MEC.

References

[1] S. Hordley. Scene illuminant estimation: past, present, and future. Color Research and Application, 31(4):303–314, 2006.

[2] Buchsbaum G. A spatial processor model for object colour perception. J Franklin Inst 1980; 310:1–26.

[3] Cardei VC, Funt B, Barnard K. Estimating the scene illuminant chromaticity by using a neural network. J Opt Soc Am A 2002; 19:2374–2386.

[4] Finlayson GD, Xu R. Convex programming colour constancy. In:Workshop on Color and Photometric Methods in Computer Vision, IEEE, October 2003. 1–7.

[5] Barnard K. Improvements to gamut mapping colour constancy algorithms In: 6th European Conference on Computer Vision, Springer, June 2000. 390–402.

[6] Finlayson GD, Hordley SD, and Hubel PM. Color by correlation: a simple, unifying framework for color constancy. IEEE Trans Pattern Anal Machine Intell 2001; 23:1209 –1221.

[7] Funt BV, Drew MS, Ho J. Color constancy from mutual reflection. Int J Comput Vis 1991; 6:5–24.

[8] K. Barnard, V. Cardei, and B. Funt. A comparison of computational color constancy algorithms; part one: Methodology and experiments with synthetic images. IEEE Transactions on Image Processing, 11(9):972–984, 2002.

[9] K. Barnard, L. Martin, A. Coath, and B. Funt. A comparison of computational color constancy algorithms; part two: Experiments with image data. IEEE Transactions on Image Processing, 11(9):985–996, 2002.

[10] S.D. Hordley and G.D. Finlayson. Re-evaluating Colour Constancy Algorithms. Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)

[11] Vlad C. Cardei and Brian Funt, "Committee-Based Colour Constancy," Proceedings of the IS&T/SID Seventh Color Imaging Conference: Color Science, Systems and Applications, 1999, pp 311-313.

[12] A. Gijsenij and T. Gevers. Color constancy using natural image statistics. In Proc. Computer Vision and Pattern Recognition, 2007.

[13] F. Tous. Computacional framework for the white point interpretation base don colour matching. Phd. Thesis, Computer Vision Center – Universitat Autònoma de Barcelona, 2006.

[14] J. Van de Weijer, C. Schmid, J. Verbeek Using High-Level Visual Information for Color Constancy, In Proc. International Conference on Computer Vision, 2007.

[15] B. Funt, K. Barnard and L. Martin. Is colour constancy good enough?. European Conference on Computer Vision, 1998, pages 445-459.

[16] Th. Gevers and A. W. M. Smeulders, Colour Based Object Recognition , Pattern Recognition (PR), 32, pp. 453-464, March, 1999

[17] D. Forsyth. A Novel Algorithm for Colour Constancy. *International Journal of Computer Vision*, 5(1):5–36, 1990.

[18] R. Benavente, M. Vanrell, and R. Baldrich. Estimation of fuzzy sets for computational colour categorization. Color Research and Application, 29(5):342{353, 2004.

[19] K. Barnard, L. Martin, B. Funt, and A. Coath, A Data Set for Colour Research. Color Research and Application, 27 (3) , pp. 147-151, 2002.

[20] Olmos, A. and F.A.A. Kingdom, A biologically inspired algorithm for the recovery of shading and reflectance images. Perception, 2004. 33: p. 1463-1473.

[21] Párraga, C.A., T. Troscianko, and D.J. Tolhurst, Spatiochromatic properties of natural images and human vision. Current Biology, 2002. 12(6): p. 483-487.

[22] P. D. Pinto, J. M. M. Linhares, and S. M. C. Nascimento, Correlated color temperature preferred by observers for illumination of artistic paintings," J. Opt. Soc. Am. A **25**, 623-630 (2008)

Biographies

Javier Vazquez received his BSc degree in Mathematics in 2006 from the Universitat de Barcelona, Spain and his MSc degree in Computer Science in 2007 from the Universitat Autònoma de Barcelona, Spain. Currently, he is a Phd. Student in the Computer Science Department and is pursuing his PhD Thesis in the colour image analysis field. He is also a researcher in the Computer Vision Center. His research interests are colour constancy and colour representation.

Maria Vanrell is an Associate Professor in the Computer Science Department of the Universitat Autonoma de Barcelona and is attached to the Computer Vision Center as a researcher. He received his Phd in Computer Science from the Unviversitat Autonoma de Barcelona in 1996. His research interest is mainly focused in colour and texture in computer vision problems, including colour constancy, texture description and colour and texture grouping.

Ramon Baldrich is an Associate Professor in the Computer Science Department of the Universitat Autonoma de Barcelona and is attached to the Computer Vision Center as a researcher. He received his Phd in Computer Science from the Unviversitat Autonoma de Barcelona in 2001. His research interest is mainly focused in colour treatment in computer vision problems, including colour segmentation, colour constancy, colour induction and image shadows.

C. Alejandro Párraga graduated in Physics (UNT, Argentina) in 1993, was awarded his MSc (Univ. of Bristol, UK) in 1996 and his PhD (Univ. of Bristol, UK) in 2003. He worked as a postdoc at the Universities. of Cambridge, and Bristol in the UK.. He was awarded both the "Juan de la Cierva"(2005) and the "Ramon y Cajal"(2007) Research Fellowships in Computer Science by the Spanish.Ministry of Science and Technology. He is based in Barcelona (Spain) since 2006