Enhancing spatio-chromatic representation with more-than-three color coding for image description

IVET RAFEGAS[1,*], JAVIER VAZQUEZ-CORRAL[2], ROBERT BENAVENTE[1], MARIA VANRELL[1], AND SUSANA ALVAREZ[3]

[1] Computer Vision Center / Computer Science Dept., Universitat Autònoma de Barcelona, Building O, Campus UAB, 08193, Cerdanyola del Vallès, Spain
[2] Dept. de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra, Roc Boronat 138, 08018, Barcelona, Spain
[3] Dept. d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Campus Sescelades, Avinguda dels Països Catalans, 26, 43007 Tarragona, Spain
[*] Corresponding author: Ivet.Rafegas@uab.cat

Compiled October 17, 2017

Extraction of spatio-chromatic features from color images is usually performed independently on each color channel. Usual 3D color spaces, such as RGB, present a high inter-channel correlation for natural images. This correlation can be reduced using color-opponent representations, but the spatial structure of regions with small color differences is not fully captured in two generic Red-Green and Blue-Yellow channels. To overcome these problems, we propose a new color coding that is adapted to the specific content of each image. Our proposal is based on two steps: (a) setting the number of channels to the number of distinctive colors we find in each image (avoiding the problem of channel correlation), and (b) building a channel representation that maximizes contrast differences within each color channel (avoiding the problem of low local contrast). We call this approach *more-than-three color coding* (MTT) to enhance the fact that the number of channels is adapted to the image content. The higher color complexity an image has, the more channels can be used to represent it. Here we select distinctive colors as the most predominant in the image, which we call color pivots, and we build the new color coding using these color pivots as a basis. To evaluate the proposed approach we measure its efficiency in an image categorization task. We show how a generic descriptor improves its performance at the description level when applied on the MTT coding. .

OCIS codes: *(010.1690) Color; (100.4994) Pattern recognition, image transforms; (100.2960) Image analysis.*

http://dx.doi.org/10.1364/ao.XX.XXXXXX

1. INTRODUCTION

In color images the values of pixels encode the spectral information of the light reflected by the surfaces in the scene. These values are represented in a k-dimensional color space, and a common formulation of this representation is written as

$$\rho_k = \int_\omega R_k(\lambda)E(\lambda)S(\lambda)d\lambda, k = 1,2,3, \quad \textbf{(1)}$$

where $E(\lambda)$ is the illuminant of the scene, $S(\lambda)$ is the surface reflectance we are looking at, $R_k(\lambda)$ is the sensitivity function of the k-th sensor defining an axis of the color space, and ω is the visible spectrum usually ranging between 400 and 700 nanometers.

Equation 1 tells us that color in the physical world is mathematically modelled as a point-based phenomenon. However, when we face the problem of solving higher level visual tasks, such as automatic image classification, to build efficient color descriptors requires the definition of color in its surrounding context.This involves the representation of spatio-chromatic information, which is a difficult problem to deal with. It has been tackled in previous works from different point of views [1–4]; here we review three main approaches.

The first one, generalized by Weickert [1], is based on considering color differences as partial derivatives computed on each RGB color channel. This was firstly addressed by Di Zenzo [5] who introduced the idea of color tensor. It provided a way of combining the channel gradients to obtain the orientation of the color variation in a local spatial neighborhood. Subsequently, this idea was further developed by Kass and Witkin [6] for oriented patterns, and finally established by Weickert [1] who introduced an additional integration scale which increases the color-spatial coherence.

A second approach by Mäenpää and Pietikäinen [2] is based on computing image descriptors in different color spaces and

using the best space for each specific application. This idea led van de Sande *et al.* [3] to study which combinations of color representation and descriptor were the most appropriate for recognition tasks. They considered well-known three-dimensional color spaces such as device-dependent RGB, colorimetric XYZ, perceptually uniform CIELab and CIELuv, cylindrical HSL and HSV, and physiologically-based opponent space. These spaces were combined with common image descriptors, such as SIFT [7] and GIST [8]. In this direction, Zhang *et al.* [9] proposed a biologically-inspired descriptor which extends the 3D color space with a fourth opponent channel. Recently, Cernadas *et al.* [10] searched for the best combination of color spaces, normalization methods and features for texture classification, and González-Rufino *et al.* [11] studied different colour-texture features to diferentiate cells in histological images.

The third approach is based on the extraction of color blobs (i.e. homogeneous color regions) directly from trichromatic representations. In particular, Alvarez and Vanrell [4] describe an image in terms of shape and color attributes of the image blobs. In this case the blobs are obtained from each channel of the opponent space by using Lindeberg's blob detector [12]. Khanina *et al.* [13, 14] adapted the scale-space technique for color images and proposed to use the Hessian matrix. Ming and Ma [15] proposed a weighted multi-scale blob detector using a hybrid operator which combines the Laplacian and the determinant of the Hessian. The results of this operator are later processed by a blob filter that includes a color-based Förstner operator and a hue-based histogram.

In all the above approaches, a variety of descriptors based on local spatial features have been defined over different three-dimensional color representations,mostly on RGB or opponent color spaces. Here, we hypothesize that the performance of these descriptors for high level visual tasks, such as image classification, can be improved by using color spaces that boost the appearance of the spatio-chromatic image structure. Boosting can be achieved by overcoming two main drawbacks: (a) inter-channel correlation of RGB spaces, and (b) lack of contrast in color-homogeneous regions of opponent spaces. These two effects can be seen in Fig. 1, where important edges between regions of different colors (orange-green edges) present clearer differences in color-opponent spaces with respect to the inter-channel correlated edges in RGB. However spatial structure appearing inside homogeneous-color regions is more contrasted in RGB than in opponent channels, where minor details (across the green or orange area) are lost.

To prove the previous hypothesis, in this paper we propose a new color representation that achieves decorrelation and enhancement of local color contrast based on the following ideas: (a) using more than three channels if required, i.e. adapting color coding to the content of each specific image; (b) enhancing local contrast inside channels by maximizing the contrast with respect to the most representative color of each channel. Following previous ideas, we compute a multi-channel representation of the spatio-chromatic image structure in a two-step process. First, we select the set of distinctive image colors, denoted as pivots, which capture the most relevant colors for each specific image. Second, the value of a pixel in each new channel is computed by the similarity between its trichromatic color and the corresponding pivot of the channel. We name the proposed representation *more-than-three* color coding, since the number of distinctive colors is not restricted to the usual three (although in some cases it can be three, or even two). In general the more color diversity the image has, the more number of color channels our representation has. We will denote our approach as MTT (*More-Than-Three*) from now on.

To test the proposed MTT coding, we use the semi-joint texton descriptor (STD) introduced by Alvarez and Vanrell [4]. This descriptor, based on the Texton theory by Julesz and Bergen [16], decomposes the image into minimal color regions (blobs). These blobs are described in terms of their color and shape attributes, which are not conditioned by the image space. This independence from the space makes this descriptor the most adequate to be directly applied to the new color representation without any additional computation. We report our results on two different experiments. Firstly, we compare the representation capabilities between MTT and two trichromatic representations, namely RGB and opponent space, concluding that MTT allows a more accurate representation of the image content thanks to the properties of presenting lower correlation and higher local contrast, that allows to get a more careful blob-based representation over the full image area. Secondly, we perform an experiment on scene categorization showing that our approach gets a higher accuracy, outperforming state-of-the art results computed at the descriptor level.

Although we show a good performance with the proposed approach, two criticisms to our initial hypothesis may arise. The first one refers to the increase in the number of color channels compared to usual representations. However, the use of extra channels can be linked to recent findings about the existence of multiple hue maps in the human visual system [17, 18]. These hue maps show selectivity to more colors than the primaries encoded in three-dimensional opponent spaces.[1] The second criticism refers to tuning to each specific image content. This tuning may complicate the description of images for comparison purposes. However, it assures obtaining a better spatio-chromatic representation for image regions that can otherwise be lost with a fixed coding as it will be shown in the experiments.

The rest of the paper is organized as follows. In Section 2 we detail our new representation. In Section 3 we define the experimental setup and present the results obtained by our approach on the experiments. Finally, in Section 4, the conclusions of the paper are discussed.

2. MORE-THAN-THREE COLOR CODING (MTT)

Our goal is to define a color representation which has a channel for each distinctive color in the image. By distinctive colors we mean those that play an important role in understanding the image content. We use as many channels as distinctive colors an image has. For a given channel we assign, (i) the maximum value to pixels of the distinctive color, and (ii) a value inversely proportional to the distance to such distinctive color to the rest of pixels. In this way, in each channel, we are maximizing the representation of a distinctive color preserving its spatial coherence. Since all the distinctive colors have their own channel, we ensure that all the important color regions of the image will be fully represented in at least one channel, and that all the region details will be maximally contrasted in its channel. We denote the distinctive color of a channel as its pivot.

Let us note here that the proposed representation is based on the content of each image. Color coding for each image is dependent on the color pivots computed from that particular image. For instance, an image of a forest with four distinctive colors could be represented by a channel for green leaves, a channel for brown tree trunks, another for blue sky, and a last one for white clouds. Meanwhile, an image of a beach could

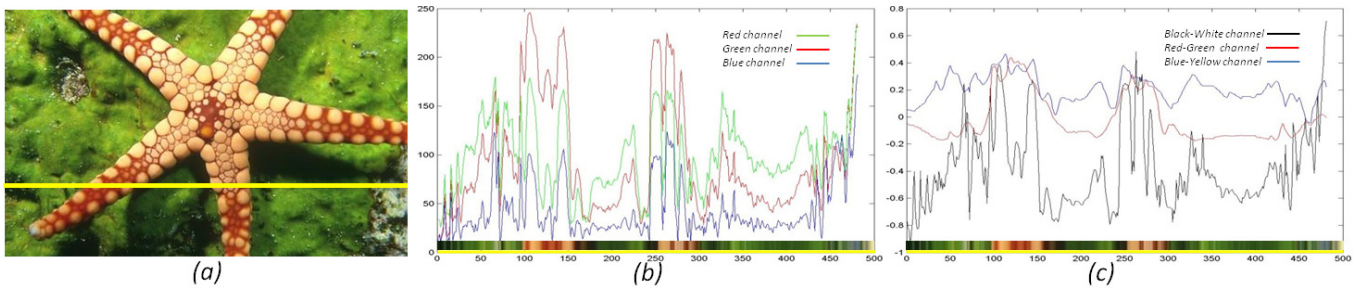


Fig. 1. Visualization of high RGB correlation and low local-contrast of color opponent channels for a single row of a natural image. (a) The analysed row is highlighted in yellow. (b) and (c) Profiles of the image row for RGB and the opponent space, respectively.

be represented by 3 channels with all the details of yellowish sand on one channel, deep blue of the sea on a second one and light blue of the sky on a third. We want to remark that this representation has not a fixed dimensionality, but it varies from one to any number representing the color complexity of a specific image scene. Nonetheless, we can state that this dimensionality usually converges to a moderate number since natural images are typically dominated by only a few colors [19].

The process to obtain the proposed MTT coding can be divided in two parts: (a) the selection of pivots (Section 2.A) and, (b) the definition of the channel values (Section 2.B). A general scheme of this process is summarized in Fig. 2.

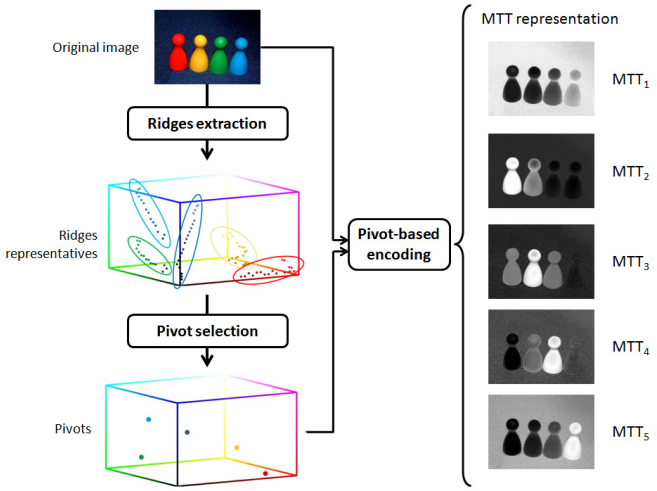


Fig. 2. Pipeline of the method. From the original image, we extract a set of ridges corresponding to the most distinctive colors and then we select a color pivot for each ridge. These color pivots are the basis to generate the proposed more-than-three (MTT) color coding of the image, which results on as many channels as distinctive colors the image has.

A. Selecting color pivots

As we have introduced before, color pivots must be the most distinctive colors of the image. In this work we propose to interpret that distinctive colors are the most predominant ones in the image and we find them using the Ridge-based Analysis of Distributions (RAD) technique [20]. The RAD algorithm groups image colors according to the ridges of the histogram. Ridges are computed by extracting all the local maxima of the histogram and connecting those which are close to each other.

Although other existing approaches could be used instead, we selected RAD because it has been proved to fulfill two properties that are of clear interest to our method. First, the RAD algorithm is invariant to some color distortions as ridges extract all the histogram maxima plus all their nearby similar values, therefore being robust to small changes like the ones caused by noise. Second, all the points in a ridge are connected, which means that the ridge representation is robust to shadows and highlights, since both shadow and non-shadow regions of an object are included in the same ridge. Thus, small color distortions will not affect our method, since they will be captured by the ridge algorithm obtaining always a single color pivot for each dominant color. These effects are not captured by classical clustering methods (e.g. k-means) which group colors mainly based on colour similarity, while RAD allows joining colors from different parts of the histogram in the same ridge, if there is a sequence of local maxima that can be connected. In the next lines we briefly summarize how predominant colors are extracted with this method.

Let us define an image I as a $M \times d$ matrix where M represents the number of pixels in the image and d is the dimension of the color space (RGB, Lab, etc.). In RAD, the first step is to look for local maxima on the color histogram $H(I)$ with the multilocal creaseness measure of Lopez *et al.* [21, 22] defined as

$$\kappa(x) = -\frac{d}{r}\sum_{k=1}^{r} \bar{\omega}^t(x_k)\cdot n(x_k), \qquad (2)$$

where x is a bin of the histogram $H(I)$, x_k is the k-th neighbor of x on an r-connected neighborhood, $\bar{\omega}(x_k)$ and $n(x_k)$ are the dominant gradient orientation and the unit normal vector to the discrete boundary of the neighborhood at each boundary site x_k, respectively, and d is the dimension of the histogram space. All the mathematical details can be found in [22]. In our implementation we use the RGB color space (i.e. $d = 3$) quantized in $30 \times 30 \times 30$ equally spaced bins. We use $r = 6$ to consider a 6-connected neighborhood as in the original implementation of RAD [20].

The local maxima of $\kappa(\cdot)$ which are close in the histogram are connected by following the lines of shallowest gradient descent until a flat region is reached. The sets of points contained in each of these lines are called ridges of the histogram and will be denoted by

$$C = \{c_1, \cdots, c_n\}, \qquad (3)$$

where c_i is a color value from the image. For a particular image, the set of all the ridges extracted applying RAD will be denoted by $\{C_i^I\}_{i=1:L}$ and they will represent the most predominant colors of the image I.

Let us now focus on searching for the color pivots. For a particular ridge C_i^I of an image I, its color pivot, ρ_i^I, is defined as the one that fulfills

$$\rho_i^I = argmax_{c \in C_i^I} H(c), \tag{4}$$

that is, ρ_i^I is the color value of ridge C_i^I that has maximum value in the image histogram $H(\cdot)$.

B. Pivot-based encoding

After selecting the set of color pivots $\{\rho_i^I\}_{i=1:L}$ of image I, we define the new spatio-chromatic representation as the $M \times L$ matrix obtained using the similarity metric given by

$$\begin{aligned} MTT_{j,i}^I &= \max_{k \in 1:M} (\|\rho_i^I - I_{k,\cdot}\|_m) - \|\rho_i^I - I_{j,\cdot}\|_m \\ &\propto 1 - \frac{\|\rho_i^I - I_{j,\cdot}\|_m}{\max_{k \in 1:M} \|\rho_i^I - I_{k,\cdot}\|_m}, \end{aligned} \tag{5}$$

where $I_{j,\cdot}$ represents the vector consisting of the 3 color components of pixel j from the original image and $\|\cdot\|_m$ represents the m-Minkowski norm. In this work we have used $m = 2$ that is equivalent to the Euclidean distance, although other distances, such as the perceptual CIEDE2000 [23], could also be used.

The computational complexity of our approach is linear in the number of pixels of the image for a fixed number of bins and a given dimension of the histogram space (in our case, $30 \times 30 \times 30$ and 3 respectively). Computing the MTT representation for an image of 768×768 pixels takes on average 888ms, from which 722ms correspond to the pivot selection (including the time of the RAD method) and 166ms to the pivot-based encoding step. These computations were done on a Intel Xeon CPU E5-1620 processor.

In Fig. 3 we present the MTT representations of a set of images, and we compare them to the RGB and the opponent representations. We can see that each MTT channel enhances different parts of the image. For example, in the first row, MTT channels emphasize different parts of the postbox. The base and the aperture are represented in the black channel, the box is in the red channel, the notice plate and the background trees are mainly enhanced on the gray channel, the grass is represented on the green channel, and the sky appears in the light-gray channel. We can appreciate how color information is less correlated on these channels than in the RGB channels (please, focus on the green and blue channels of RGB) and that opponent channels present less contrast between the different objects of the image. Similarly, in the second row, the different parts of the boy's clothes (in red, blue, and orange channels), the snowman (in the white channel), and the background (in the gray channel) are all enhanced in different channels. An analogous analysis can be performed in the rest of images.

Notice that since our MTT representation is content-based we obtain a different number of channels on each image depending on the variety of colors in it. In the examples, the first two images have five channels whereas the last one showing a purple flower on a green background has only two channels. Notice also that the MTT channels represent different colors for each image. In some cases, as in the first-row image, two shades of the same color can be represented in different channels if they are sufficiently different from each other (in this example, gray and light gray).

Finally, let us explain how we can derive an inverse transform to the original space. By the construction of our space we know that for each channel: i) the color selected as a pivot is always a trichromatic value appearing in the image. Therefore, the maximum value of the channel is equal to the maximum difference between the color of the pivot and the color of a certain pixel in the image; and ii) there exists a pixel in the image (the one with its color at a further distance of the pivot) whose representation in the channel is 0. Mathematically,

$$\max_{k \in 1:M} MTT_{k,i}^I = \max_{k \in 1:M} \|\rho_i - I_{k,\cdot}\|_m. \tag{6}$$

$$min_{k \in 1:M} MTT_{k,i}^I = 0. \tag{7}$$

These two properties, allow us to invert Eq.5 as follows

$$\|\rho_i^I - I_{j,\cdot}\|_m = \max_{k \in 1:M} MTT_{k,i}^I - MTT_{j,i}^I. \tag{8}$$

Then, given $MTT_{j,i}^I$ and ρ_i this last equation defines a surface (an sphere if $m = 2$) of possible values for each $I_{j,\cdot}$. Therefore, to recover the original image we just need to know the value of three of the pivots that are linearly independent, and use trilateration. Then, our recovered image will be given by values $I_{j,\cdot}$, that fulfill Eq. 8 for three values of i.

C. Illumination invariance

As explained in the introduction, pixel values of an image depend on the reflectance of the objects, the camera sensors, and the illumination of the scene. Therefore, when the illumination of the scene changes (which is usual in real images), pixel values also change thus hindering the performance of computer vision algorithms. Different methods have been proposed to counteffect the illuminant variability, either by discounting the illuminant [24] or by performing some form of color normalization [25]. In this section, we show that our image representation can be directly used as an invariant to the illumination (therefore avoiding the need of further processing) when computed on the logRGB color space.

In RGB space, the change in illumination between two images of the same scene can be approximately modeled by a single scaling factor on each channel (i.e. the Von Kries coefficient law [26]), either directly [27] or by applying the spectral sharpening technique [28, 29]. This is, given an image I^1, an image I^2 of the same scene under a different illuminant can be defined as

$$I^2 = \mathcal{D}^{1,2} I^1, \tag{9}$$

where $\mathcal{D}^{1,2}$ is a 3×3 diagonal matrix containing the scaling factors for each RGB channel, therefore transforming the colors under the first illuminant to those under the second illuminant. If we apply a logarithm operation to the RGB space, the previous equation can be rewritten as

$$\log(I^2) = \left[d^{1,2}, \cdots, d^{1,2}\right] + \log(I^1), \tag{10}$$

where $d^{1,2} = \left[\log(D_{11}^{1,2}), \log(D_{22}^{1,2}), \log(D_{33}^{1,2})\right]^T$. This is the case since $\mathcal{D}^{1,2}$ is a diagonal matrix and thus the channels of I^1 are treated independently. Equation 10 tells us that an illumination change can be modeled by a translation in logRGB space. Therefore, for any color value $x \in$ logRGB we have that

$$H^2(x) = H^1(x + d^{1,2}), \tag{11}$$

where $H^1(\cdot)$ and $H^2(\cdot)$ denote the histograms of $\log(I^1)$ and $\log(I^2)$ respectively. Consequently, following Section 2.A, we

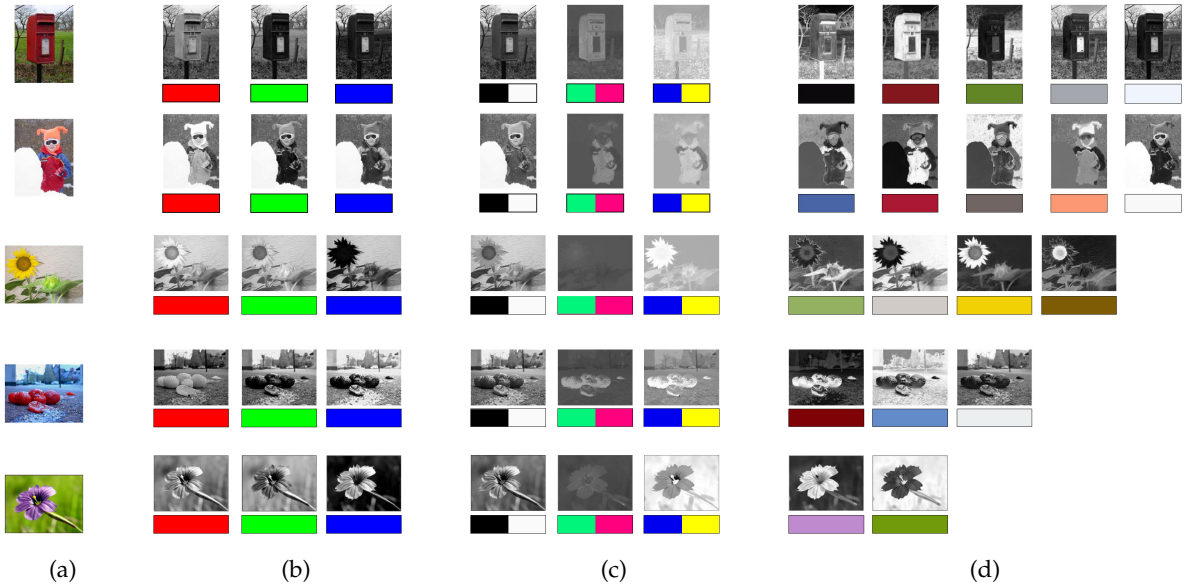


Fig. 3. Examples of MTT representation for several images and comparison to the RGB and opponent representations. (a) Original images. (b) RGB channels. (c) Opponent channels. (d) MTT channels. On channel images, values are represented on grayscale (black=0, white=1). The color boxes under channel images show the correspondence with RGB, opponent, and MTT channels. The proposed MTT represents images with a variable number of channels depending on the number of distinctive colors the image has.

have that the color pivots of $\log(I^1)$ and $\log(I^2)$ are also related by

$$\rho_i^{\log(I^2)} = \rho_i^{\log(I^1)} + d^{1,2}. \tag{12}$$

From Eq. 12 and Eq. 5 we have

$$\begin{aligned} MTT_{j,i}^{\log(I^1)} = \\ \max_{k\in 1:M}(\|\rho_i^{\log(I^1)} - \log(I^1_{k,\cdot})\|_m) - \|\rho_i^{\log(I^1)} - \log(I^1_{j,\cdot}))\|_m = \\ \max_{k\in 1:M}(\|\rho_i^{\log(I^2)} - d^{1,2} - (\log(I^2_{k,\cdot}) - d^{1,2})\|_m) - \\ \|\rho_i^{\log(I^2)} - d^{1,2} - (\log(I^2_{j,\cdot}) - d^{1,2})\|_m = \\ \max_{k\in 1:M}(\|\rho_i^{\log(I^2)} - \log(I^2_{k,\cdot})\|_m) - \|\rho_i^{\log(I^2)} - \log(I^2_{j,\cdot})\|_m = \\ = MTT_{j,i}^{\log(I^2)}. \end{aligned} \tag{13}$$

Therefore, our representation computed on logRGB space is approximately invariant to the illuminant. An example of this invariance is shown in Fig. 4, where we can see, from left to right, the original RGB image, the results of the MTT representation fixing the number of channels to 3, and a visualization of the MTT channels concatenated as an RGB image. It is clear that the MTT channels are very similar for all the images, making the RGB-like visualization stable under illuminant changes.

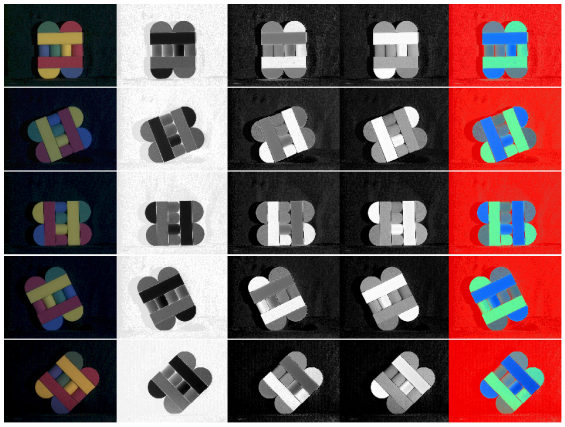


Fig. 4. Example of the approximate illuminant invariance of our representation. From left to right: original RGB image, the three MTT channels, and the MTT representation visualized as a RGB image. Illuminants in the original RGB images are (from top to bottom) an approximate D65 illuminant, an illuminant with $CCT = 3500$K plus a blue filter, an illuminant with $CCT = 4700$K, an approximate D65 illuminant plus a blue filter, and a fluorescent light.

3. EXPERIMENTS AND RESULTS

As presented in the previous section, MTT provides a new color representation which is based on the specific content of each image. In this section we show its power to build generic color image descriptors. The evaluation is performed in two steps. We firstly evaluate how MTT overcomes the problems of RGB and opponent spaces to encode spatio-chromatic information of images. We evaluate this improvement in terms of the channels correlation and local contrast, and also showing how MTT representation improves the ability of an specific image descriptor, and secondly we evaluate how MTT increases the performance of a descriptor in a scene classification task.

The evaluation is performed in two steps. We firstly evaluate how MTT overcomes the problems of RGB and opponent spaces

to encode spatio-chromatic information of images. We evaluate this improvement in terms of the channels correlation and local contrast, and also showing how MTT representation improves the ability of an specific image descriptor.

Considering the problem of generic image description, the comparison between descriptors of different images built on the MTT representation requires to be adapted to any number of channels. To overcome this problem we use a the Semi-Joint Texton descriptor (STD) [4] and a variant of it, both are explained on the next subsection. This descriptor gives an intermediate-level representation in terms of image blobs, i.e. color-homogeneous convex regions, that is computed regardless of the color space.

Taking into account the previous considerations, we organize this experimental section in four subsections. Firstly, we introduce the image descriptor used in the experiments. Secondly, we provide the details of the setup used in the experiments, which are fully explained in the remaining two subsections.

A. Image description: Semi-joint Texton Descriptor

The Semi-joint Texton Descriptor (STD) introduced by Alvarez and Vanrell in [4] describes an image in terms of shape and color attributes of the image blobs. STD can be computed on any color space and we show that the performance of this descriptor on scene recognition is improved when MTT is used instead of RGB or the opponent representation. An interesting property of this descriptor is that the attributes of the blobs it uses do not depend on the input color space where the blobs are initially detected. Due to this property, the descriptions of two images can be compared independently of the color representation where the blob detection is performed, even if their representations have different number of channels.

The STD algorithm starts detecting the blobs of an image by applying a multi-scale Laplacian in each separate channel of the image representation of choice. From the blobs detected in all the channels, color and shape attributes are extracted. Then the STD is defined as a combination of shape (STD_S) and color (STD_C) descriptors of image blob's attributes (see sections 3.A.1 and 3.A.2):

$$STD = [STD_S \; STD_C]. \tag{14}$$

A.1. Shape descriptor

The shape descriptor is a histogram of blobs' shape attributes. For each detected blob, shape attributes, namely area, orientation, and aspect ratio, are obtained independently of the color channel where the blob was detected. Then, all blobs' attributes are quantized in a three-dimensional blob-shape space in order to compute the histogram. In this histogram each bin represents a visual word of the universal shape vocabulary defined by the quantization of the blob-shape space.

A.2. Color descriptor

The color descriptor is a histogram of blobs' color attributes. The histogram is computed in the HSI color space, where blobs' color attributes are quantized. In this histogram each bin represents a visual word of the universal color vocabulary defined by the quantization of the color space (see figure 9 in [4]).

In this paper, we also use a variant of the color descriptor defined in [30]. This approach is based on the color-naming model of Benavente *et al.* [31], which categorizes any image pixel p in one of the 11 basic colors defined by Berlin and Kay [32] (i.e. red, green, blue, yellow, orange, brown, pink, purple, white, gray, and black). Such categorization is done by means of an 11-dimensional membership vector $\mu(p)$, where each component $\mu_i(p)$ can be interpreted as the probability of color p to belong to a particular color $\mathcal{C}_i$. Pixels are assigned to the color term with highest membership, which is then backed up with a modifier related to its lightness (i.e. dark, medium, or light). Using this color-naming representation the quantization of the color space is more perceptual than the original quantization [4], where just an equally-spaced division of the space was used.

To avoid confusions, from now on we denote by STD_{OR} the original descriptor defined in [4] (shape descriptor plus color descriptor on HSI), and by STD_{CN} the variant which uses color naming for the color description [30] (i.e. STD_{CN} is formed by the shape descriptor and the color descriptor based on color names). Figure 5 shows a graphical representation of the two STD implementations used in this work.

A.3. Adding spatial layout information

The STD descriptor is a global first-order statistic of blob attributes. For scene recognition the insertion of the spatial layout is a must since similar color areas can represent different things depending on their location in the image. For example, medium and large blue blobs can represent either water (e.g. a lake or the sea) or sky; in this sense, adding their spatial location will help to distinguish if they represent water (usually located at the bottom images) or sky (usually located at the top).

Hence, we add the spatial component similarly to how it is added in the GIST descriptor [8]. Given an image I we decompose it in a set of non-overlapping sub-images $I_1, \cdots, I_k$, which are obtained by dividing each of the image dimensions by a particular natural number (usually 2, 3, or 4). Then, we compute the descriptor for each of the sub-images and concatenate them, obtaining a final descriptor of the form

$$STD = [STD_{S_1} \cdots STD_{S_k} \; STD_{C_1} \cdots STD_{C_k}], \tag{15}$$

where STD_{S_i} and STD_{C_i} represent the shape and color descriptors of sub-image I_i.

B. Experimental setup

In our experiments, the maximum number of channels for the MTT representation is set to $L = 8$. This value was experimentally found by testing values from $L = 2$ to $L = 11$. Results gradually improve as the value of L increases, but for $L > 8$ the improvement is not significant. In case that more than 8 ridges are extracted from an image (see Section 2.A), the 8 ridges that represent the largest areas of the image (computed via a watershed in the color histogram of the image) are selected.

To obtain the shape descriptor, we use the following quantization of the shape space: 8 orientations ($0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°$), 7 scales (area), and 3 aspect ratio values (isotropic, elliptical, and highly-elongated). Isotropic blobs are assigned to orientation $0°$. Thus the shape descriptor has dimension $119 = (8 \text{ orientations} \times 7 \text{ scales} \times 2 \text{ aspect ratios}) + 7$ (one bin per scale for isotropic blobs).

In the case of the color descriptor we have used the two configurations explained in Section 3.A.2. For STD_{OR}, the HSI color space is quantized in 16 bins for H, 4 for S, and 5 for I, making a size of the color descriptor of 320 bins. For STD_{CN}, color is defined in terms of 11 names and 3 modifiers, which gives a size of 33 bins for the color descriptor. Therefore, the total size of STD_{OR} is $119 + 320 = 439$ bins, whereas STD_{CN} has a the total size is $119 + 33 = 152$ bins. If spatial decomposition is used (see Section 3.A.3), these values should be multiplied by

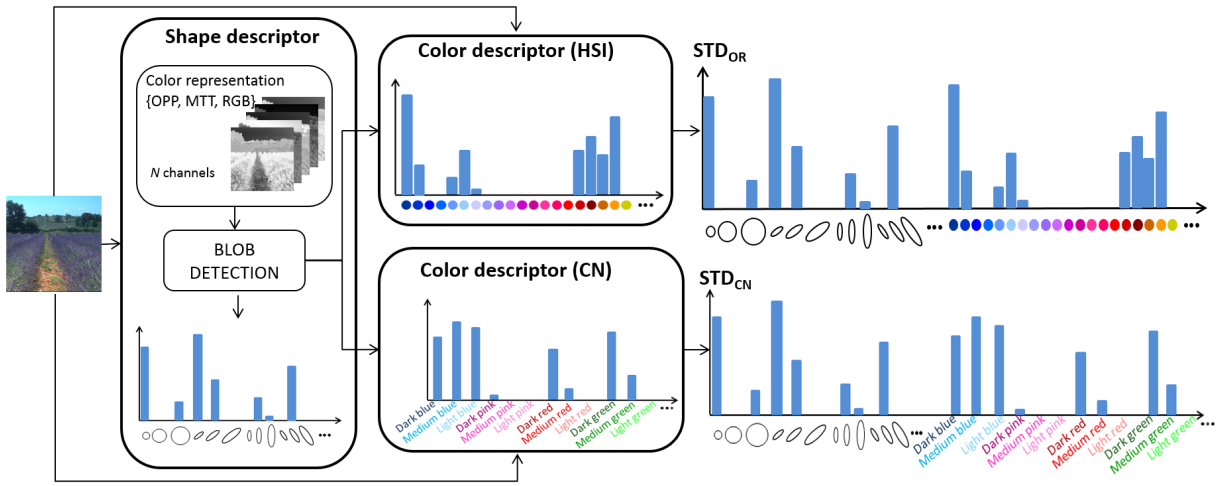


Fig. 5. Diagram of the process to obtain STD_{OR} [4] and STD_{CN}[30]. Blobs are detected on each channel of the chosen color representation and shape attributes are computed to generate the shape descriptor STD_S. The color descriptor STD_C is computed either on the HSI color space or using color names.

the number of sub-images considered to obtain the final size of the descriptor.

Finally, the dataset used in all the experiments is the dataset of scenes created by Oliva and Torralba [8], which contains 2688 images of 256 × 256 pixels from 8 categories: coast, forest, highway, inside city, mountain, open country, street and tall building.

C. Experiment 1: Analysis of MTT properties

In this first experiment we analyze the properties of the proposed color representation. As we mentioned in the introduction, the main problems of usual color spaces to encode the spatio-chromatic image structure are due to the high correlation between channel and the lack of local contrast for specific colors. These two properties are inherent to the channel-based representation derived from the sensor that reduces the capability to represent all the image details. Even when we transform to an opponent representation, the lack of contrast of the new chromaticity channels does not allow representing all the details of areas with homogeneous chromaticity. Considering these two aspects, in this experiment we have computed the inter-channel correlation and the channel's local contrast for RGB, normalized opponent space [2] (nOPP), and the MTT representation. have also considered the space defined by the three eigenvectors obtained by PCA on the RGB space.

For a given image, the inter-channel correlation has been computed as the average of the minimum pairwise-channel correlation[3] obtain the local contrast, we use the method defined by Haun and Peli [33].

The results are shown in Table 1. We can see that the MTT representation presents a combined result of low inter-channel correlation and high local contrast. If these results are compared to the ones obtained by the opponent space, we see that MTT obtains better results in both measures. PCA presents the lowest correlation at the cost of also obtaining the lowest local contrast. Comparing to the RGB space, local contrast of RGB channels is slightly higher than in MTT, but in RGB the correlation between its channels is considerably higher than in MTT. We also looked at the behavior of local contrast when considering only the three

Table 1. Correlation among the different channels and mean local contrast for the different color spaces for all the images on the Oliva and Torralba dataset.

	Correlation	Local contrast
RGB	0.82 (± 0.18)	20.47 (± 10.21)
nOPP	0.30 (± 0.20)	10.75 (± 7.48)
PCA	0.00 (± 0.02)	9.55 (± 8.26)
MTT	0.25 (± 0.18)	19.24 (± 10.76)

MTT channels that have higher local contrast for each image. In this case, the result for MTT is over 10% higher than in RGB, therefore showing that a subset of the MTT channels presents higher local correlation than any other representation of the same dimension.

Let us now show how the better results of MTT in correlation and local contrast allow us for a better image description. To this end, we detect the blobs in each image of the dataset (using the blob descriptor encoded in the STD descriptor) on different color representations to analyze how well these blobs describe the content of the image. We assume that, in general, the most area covered by detected blobs, the best the overall appearance of the image will be described. Thus, an image can be reconstructed by plotting their blobs at the locations where they were detected, and filling them with their color attribute. Figure 6 shows a visual comparison between the blobs detected on the proposed MTT, the normalized opponent color space, and the RGB space. We can appreciate that on MTT more parts of the image are described, the details are better represented and the overall structure of the original image (i.e. the gist of the image) is more appreciable.

To give a quantitative analysis of the results in the previous figure, in Table 2 we show the percentages of covered area by blobs detected on RGB, the opponent space and MTT. As it can be seen, the percentage of area covered by blobs detected

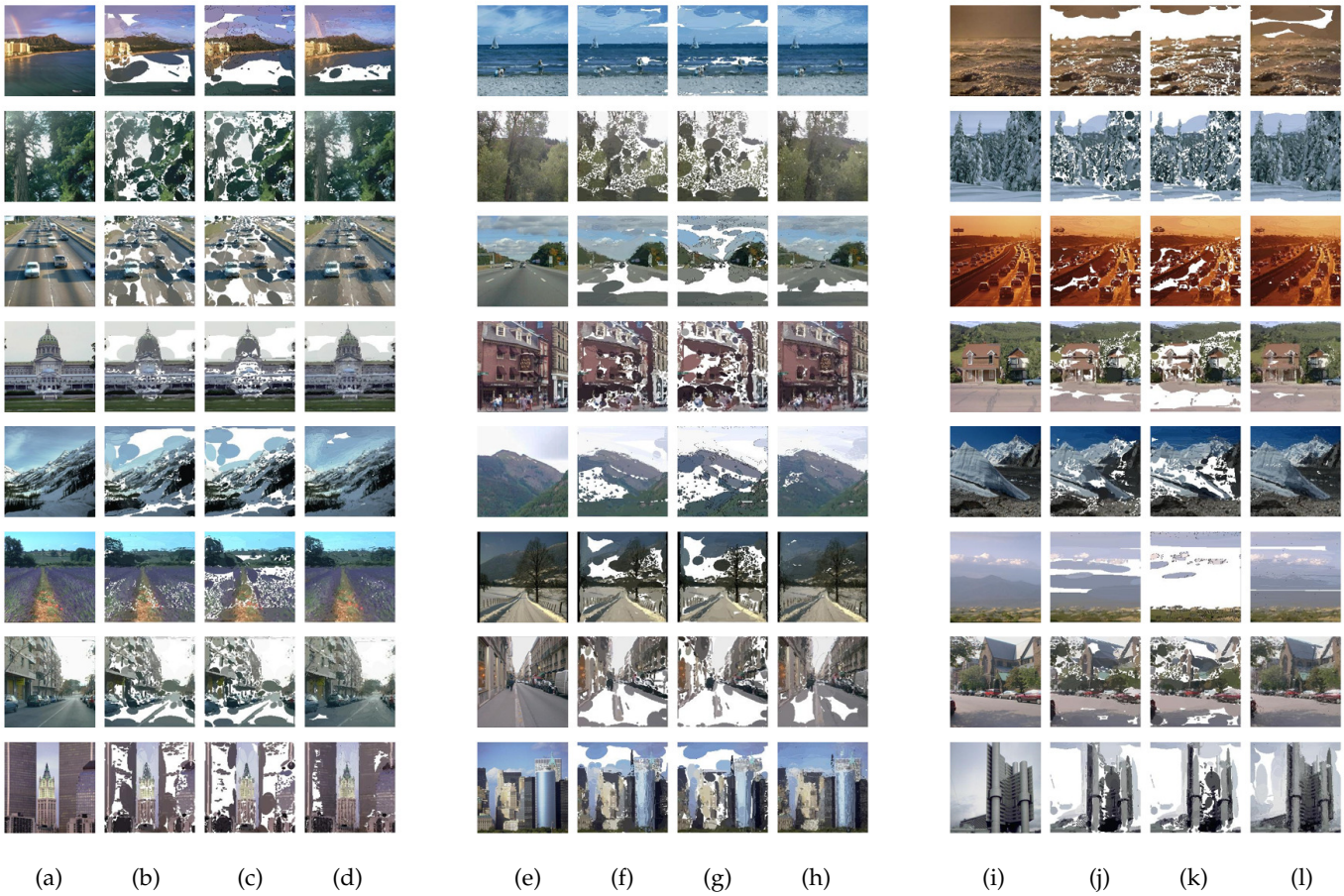


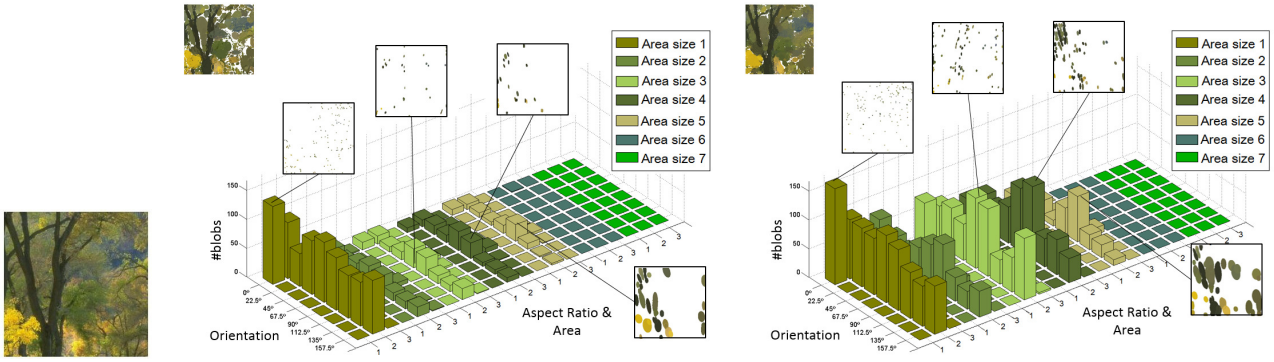
Fig. 6. Blobs detected on different color representations. Each row corresponds to one of the categories of the dataset. (a), (e) and (i) Original images. (b), (f) and (j) Blobs detected using the RGB color space. (c), (g) and (k) Blobs detected using the normalized opponent space. (d), (h) and (l) Blobs detected using the MTT representation.

on MTT is higher than the ones obtained on the other color representations. This increase can be found in all the categories of the dataset. For example, in the forest category the increase is over 13% with respect to RGB. This could be due to the fact that images from this category have low contrast and similar hues, which makes that areas of similar color can not be detected as different regions in the opponent or the RGB channels. By contrast, MTT is more able to represent different shades of the same hue in different channels which facilitates the posterior blob detection.

Finally, let us analyze how our better detection of the gist of the image translates to the shape descriptor part STD_S. To this end, in Fig. 7 we compare the distributions of detected blobs from an image using opponent and MTT representations. Distributions are displayed as 3D histograms where one of the axes represents orientation, another represents aspect ratio and area jointly, and the third represents the number of blobs. We can appreciate that each visual word in STD_S clusters blobs with a similar visual appearance (i.e similar area, orientation, and aspect ratio). We note that STD_S on the MTT channels detects more blobs than on the opponent space, specially on those bins where some blobs are already detected on the opponent space. Moreover, MTT allows to detect blobs with attributes corresponding to bins where only a few blobs are detected on the opponent channels. These extra blobs detected on the MTT representation are mainly found at large uniform areas, which explains why MTT is more effective representing the overall structures of the image as we have seen in Figure 6.

Table 2. Percentages of covered area for each category on the Oliva and Torralba dataset using STD_{OR} descriptor on RGB, normalized opponent (nOPP) and MTT channels.

Category	RGB	nOPP	MTT
Coast	85.52%	78.41%	96.10%
Forest	84.61%	83.36%	97.95%
Highway	79.32%	69.03%	91.52%
Inside city	90.01%	85.06%	98.75%
Mountain	85.28%	82.44%	96.24%
Open country	90.93%	87.30%	97.71%
Street	81.74%	73.94%	95.06%
Tall building	87.33%	83.65%	96.43%
All	**85.94%**	**80.99%**	**96.38%**



(a) (b) (c)

Fig. 7. Comparison of the shape descriptors of an image (a) obtained on the opponent space (b) and on the MTT representation (c). Shape descriptors are shown as a 3D histogram where each bin clusters detected blobs with similar area, orientation and aspect ratio. On the aspect ratio and area axis, '1' corresponds to isotropic blobs, '2' to elliptical blobs, and '3' to highly-elongated blobs. Bins corresponding to different areas are plotted in different colors. Area increases along the axis.

D. Experiment 2: Scene recognition

In this experiment we test the efficiency of the new representation when it is used to compute the STD for scene recognition tasks. We first compare different spatial decompositions to determine the best configuration of STD and then we compare the results to the state of the art on the database of Oliva and Torralba [8]. The experiments are done following the same methodology used in [34]. A linear support vector machine is trained and tested on a randomly selected split of 600 images for training and 120 images for testing. This procedure is repeated 10 times and results are averaged.

D.1. Analysis of spatial decomposition

As stated in Section 3.A.3, the inclusion of spatial information on STD can improve its results for general tasks in computer vision. Spatial information is a building part in some image descriptors, such as GIST [8], but it should not be confused with the idea of spatial pyramids [35], where the descriptor is computed on regions of different sizes and are later combined into a single descriptor.

To analyze the relationship between the number of sub-images used and the accuracy achieved, we have computed the results of the original implementation of STD (STD_{OR}) and STD using color names (STD_{CN}) on different color spaces, and considering the whole image (no spatial decomposition) and different number of sub-images (4, 9, and 16). According to these results (see Fig. 8), the inclusion of spatial information in the descriptor by dividing the image into 4 sub-images increases the accuracy by at least 4% in all cases. Considering 9 sub-images still increases the accuracy but the increase is not as remarkable as in the previous case. After that, the increase is not significant or there is even a slight decrease in accuracy in the case of the descriptors computed on RGB.

D.2. Comparison to state of the art

Given the results of the previous section, we use 4 sub-images to compute STD_{OR} and STD_{CN} because this configuration provides us with good performance and the size of the descriptor does not increase dramatically (1756 for STD_{OR} and 608 for STD_{CN}). Now, these results are compared to the ones reported in [34] for three well-known descriptors: SIFT [7], GIST [8], and

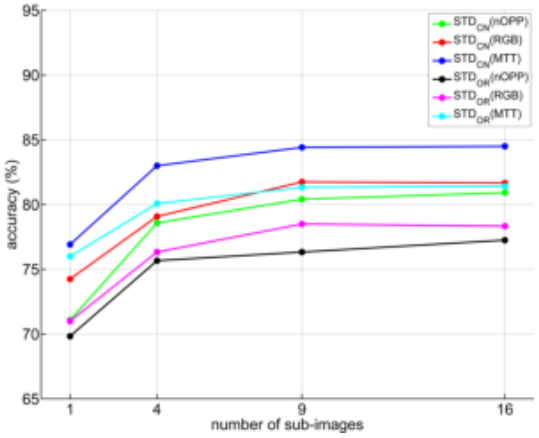


Fig. 8. Accuracy on scene recognition in terms of the number of sub-images used to compute STD descriptors on different color spaces.

HMAX [36] and are presented in Table 3. Rows 1 to 3 summarize the results of Brown and Süsstrunk in [34]. We only report the color space where each descriptor achieved the best results. The highest accuracy was obtained with GIST on the opponent space (without normalization). Rows 4 to 6 and 7 to 9 show the performance of STD_{OR} and STD_{CN}, respectively. In both cases, the descriptor is computed on three color representations (RGB, normalized opponent space, and MTT).

Analyzing the results, the use of MTT on both STD descriptors provides with an improvement on the accuracy of about 4% and 4.5% comparing to RGB and the normalized opponent space, respectively. This result can also be observed in Fig. 8 where for any number of sub-images, any descriptor computed on MTT overcomes the same descriptor computed on RGB or on the normalized opponent space.

Moreover, we computed the Wilcoxon test with the hypothesis that the results obtained with GIST on the opponent space and with STD_{CN} on MTT in the 10 trials of the experiments belonged to the same distribution. We obtained a p-value of 0.0020 with a significance level of 5%. Therefore, we can reject our

Table 3. Accuracy (%) and standard deviation computed over 10 trials on the scene recognition experiment. Results for HMAX, GIST, and SIFT were extracted from [34]. In parenthesis we show the color space used.

Descriptor	Accuracy (%)
SIFT(nOPP)	69.6 ($\pm$ 2.5)
HMAX(RGB)	74.0 ($\pm$ 4.4)
GIST(OPP)	77.8 ($\pm$ 3.4)
STD_{OR}(nOPP)	75.7 ($\pm$ 3.5)
STD_{OR}(RGB)	76.3 ($\pm$ 3.5)
STD_{OR}(MTT)	**80.1** ($\pm$ 3.6)
STD_{CN}(nOPP)	78.6 ($\pm$ 2.5)
STD_{CN}(RGB)	79.1 ($\pm$ 4.6)
STD_{CN}(MTT)	**83.0** ($\pm$ 3.0)

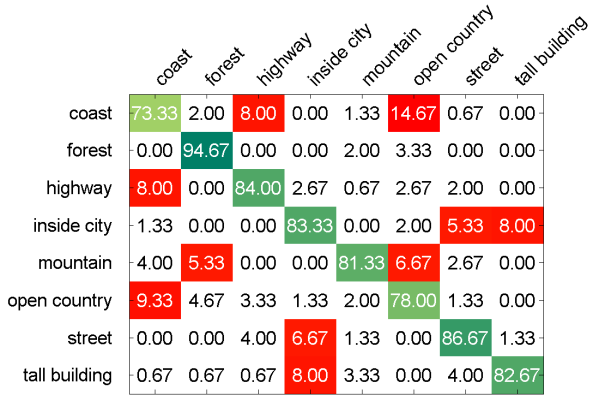


Fig. 9. Mean confusion matrix of the scene recognition experiment on Oliva and Torralba dataset using STD_{CN} computed on the MTT representation. Greenish cells correspond to the results with accuracy higher than 70%, while reddish cells correspond to the misclassified results having a percentage above 5%.

null hypothesis and conclude that the improvement obtained by STD_{CN} on MTT over GIST on the opponent space are statistically significant. Therefore, we can conclude that the use of MTT improves the results of the STD descriptors with respect to the use of RGB or the normalized opponent space. Furthermore, both STD_{OR} and STD_{CN} computed on MTT outperform GIST results reported in [34]. Let us remark here that our best result (STD_{CN} on MTT channels with an accuracy of 83.0%) is obtained with a descriptor composed by 608 bins, while the GIST descriptor has a size of 960 bins.

Moving to the analysis by category, Fig. 9 shows the confusion matrix of our best result (STD_{CN} on MTT). Each cell of the matrix shows the percentage of images of a class (row) classified as each of the classes (columns). From the matrix we can see that the category with higher accuracy is forest. This could be expected since this category shows a low intraclass variability. By contrast, open country and coast present a high confusion (*e.g.* 14% of coast images are classified as open country). Similarly, city and tall building are two categories with a certain confusion (8% of images of each class classified in the other class). Both cases can be explained by the fact that images in these pairs of categories show high similarities; for example, open country has many images of lakes and rivers that can be confused with images of coast, and city category contains many images of buildings combined with other elements such as cars and pedestrians that can be confused with images from the tall building category.

4. CONCLUSIONS

The main novelty of this work is the creation of a new color representation based on the specific content of the image. With this approach we aim an image color coding that enhances spatio-chromatic information and reduces inter-channel correlation. The goal is achieved in a two-step process. Firstly, we set the number of channels used in MTT with the number of relevant colors the image has, defined as pivots. Secondly, we build individual channel representation that maximizes contrast differences using a similarity metric with respect to the color pivot related to each channel.

The proposed approach presents some clear advantages:

- Represents images according to its own color complexity, this is with more than three dimensions if required. As each dominant color is mostly represented in one of the dimensions, our approach shows more ability to capture the image details.
- Increases the local contrast and reduces the correlation of the resulting channels, which plays a crucial role in several tasks such as edge and blob detection, segmentation, and recognition.
- Presents illuminant invariance properties if it is built onto a log space. This can be an important benefit, essentially in recognition tasks.
- Increases performance when applied to build color image description for scene classification. This increase is mainly due to the improvement in the blob detection step of the color descriptor.

To prove these advantages we have performed two experiments. First, a qualitative experiment to show the performance of the MTT representation in a blob detection task. We visualize how the proposed approach presents low correlation and high local contrast, and how it improves the area covered by detected features across the full image plane. A second quantitative experiment has been performed for scene recognition. We show how the same descriptor improves its performance when applied on the MTT coding, and we compare the results to current state-of-art descriptors, which are overcome.

In the future, we plan to study the impact of MTT to detect keypoints. Many descriptors use the Harris-Laplace detector to select keypoints in the image where the descriptor is computed. The increase in the image contrast in the MTT channels could allow the Harris-Laplace operator to detect more points and this fact could improve the results of the image descriptors computed on those locations.

FUNDING INFORMATION

The CERCA Programme of the Generalitat de Catalunya; Spanish Ministry of Economy and Competitiveness (TIN2014-61068-

R, TIN2015-71537-P, IJCI-2014-19516); European Research Council (Starting Grant 306337).

NOTES

[1]Hue maps are defined as clusters of neurons that peak when a specific color stimuli is presented. Although, a lot of research is left to be done in this area, some interesting results have started to arise: there are more hue maps in higher levels than the six opponent colors [18, 37], and the peaks of the cell responses are given by particular hues [38–40].

[2]As defined in the C-SIFT descriptor [3].

[3]We use this measure instead of a global correlation average due to the different number of channels in each color representation.

REFERENCES

1. J. Weickert, "Coherence-enhancing diffusion of colour images." Image and Vision Computing **17**, 201–212 (1999).
2. T. Mäenpää and M. Pietikäinen, "Classification with color and texture: jointly or separately?" Pattern Recognition **37**, 1629–1640 (2004).
3. K. van de Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence **32**, 1582–1596 (2010).
4. S. Alvarez and M. Vanrell, "Texton theory revisited: a bag-of-words approach to combine textons," Pattern Recognition **45**, 4312–4325 (2012).
5. S. Di Zenzo, "A note on the gradient of a multi-image," Computer Vision, Graphics, and Image Processing **33**, 116–125 (1986).
6. M. Kass and A. Witkin, "Analyzing oriented patterns," Computer Vision, Graphics, and Image Processing **37**, 362–385 (1987).
7. D. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision **60**, 91–110 (2004).
8. A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International Journal of Computer Vision **42**, 145–175 (2001).
9. J. Zhang, Y. Barhomi, and T. Serre, "A new biologically inspired color image despcriptor," in "12th European Conference on Computer Vision," (2012), pp. 312–324.
10. E. Cernadas, M. Fernández-Delgado, E. González-Rufino, and P. Carrión, "Influence of normalization and color space to color texture classification," Pattern Recognition **61**, 120–138 (2017).
11. E. González-Rufino, P. Carrión, E. Cernadas, M. Fernández-Delgado, and R. Domínguez-Petit, "Exhaustive comparison of colour texture features and classification methods to discriminate cells categories in histological images of fish ovary," Pattern Recognition **46**, 2391–2407 (2013).
12. T. Lindeberg and J. Eklundh, "On the computation of a scale-space primal sketch," Journal of Visual Communication and Image Representation **2**, 55–78 (1991).
13. N. Khanina, E. Semeikina, and D. Yurin, "Color blob and line detection in scale-space," Pattern Recognition and Image Analysis **21**, 267–269 (2011).
14. N. Khanina, E. Semeikina, and D. Yurin, "Scale-space color blob and ridge detection," Pattern Recognition and Image Analysis **22**, 221–227 (2012).
15. A. Ming and H. Ma, "A blob detector in color images," in "6th ACM International Conference on Image and Video Retrieval," (2007), CIVR'07, pp. 364–370.
16. B. Julesz and J. Bergen, "Textons, the fundamental elements in preattentive vision and perception of textures," Bell System Technical Journal **62**, 1619–1645 (1983).
17. H. Tanigawa, H. Lu, and A. Roe, "Functional organization for color and orientation in macaque V4," Nature neuroscience **13**, 1542–1548 (2010).
18. L. Parkes, J. Marsman, D. Oxley, J. Goulermas, and S. Wuerger, "Multivoxel fMRI analysis of color tuning in human primary visual cortex," Journal of Vision **9**, 1,1–13 (2009).
19. I. Omer and M. Werman, "Color lines: Image specific color representation," in "IEEE Conference on Computer Vision and Pattern Recognition," (2004), pp. 946–953.
20. E. Vazquez, R. Baldrich, J. van de Weijer, and M. Vanrell, "Describing reflectances for colour segmentation robust to shadows, highlights and textures," IEEE Transactions on Pattern Analysis and Machine Intelligence **33**, 917–930 (2011).
21. A. Lopez, F. Lumbreras, J. Serrat, and J. Villanueva, "Evaluation of methods for ridge and valley detection," IEEE Transactions on Pattern Analysis and Machine Intelligence **21**, 327–335 (1999).
22. A. M. Lopez, D. Lloret, J. Serrat, and J. J. Villanueva, "Multilocal creaseness on the level-set extrinsic curvature," Computer Vision and Image Understanding **77**, 111–144 (2000).
23. M. R. Luo, G. Cui, and B. Rigg, "The development of the cie 2000 colour-difference formula: Ciede2000," Color Research & Application **26**, 340–350 (2001).
24. J. T. Barron, "Convolutional color constancy," in "IEEE International Conference on Computer Vision," (2015), pp. 379–387.
25. G. Finlayson and R. Xu, "Illuminant and gamma comprehensive normalisation in logRGB space," Pattern Recognition Letters **24**, 1679 – 1690 (2003).
26. G. Wyszecki and W. Stiles, *Color science: concepts and methods, quantitative data and formulae* (John Wiley & Sons, 1982), 2nd ed.
27. G. Finlayson, M. Drew, and B. Funt, "Color constancy: Generalized diagonal transforms suffice," Journal of the Optical Society of America A **11**, 3011–3020 (1994).
28. G. Finlayson, M. Drew, and B. Funt, "Spectral sharpening: sensor transformations for improved color constancy," Journal of the Optical Society of America A **11**, 1553–1563 (1994).
29. J. Vazquez-Corral and M. Bertalmío, "Spectral sharpening of color sensors: Diagonal color constancy and beyond," Sensors **14**, 3965–3985 (2014).
30. S. Alvarez, "Revisión de la teoría de los textons. enfoque computacional en color," Ph.D. thesis, Universitat Autònoma de Barcelona (2010).
31. R. Benavente, M. Vanrell, and R. Baldrich, "Parametric fuzzy sets for automatic color naming," Journal of the Optical Society of America A **25**, 2582–2593 (2008).
32. B. Berlin and P. Kay, *Basic Color Terms: Their Universality and Evolution* (University of California Press, Berkeley, CA, 1969).
33. A. Haun and E. Peli, "Perceived contrast in complex images," Journal of Vision **13**, 3,1–21 (2013).
34. M. Brown and S. Süsstrunk, "Multispectral SIFT for scene category recognition," in "IEEE Conference on Computer Vision and Pattern Recognition," (2011), pp. 177–184.
35. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for rcognizing natural scene categories," in "IEEE Conference on Computer Vision and Pattern Recognition," (2006), pp. 2169–2178.
36. J. Mutch and D. Lowe, "Object class recognition and localization using sparse features with limited receptive fields," International Journal of Computer Vision **80**, 45–57 (2008).
37. M. Li, F. Liu, M. Juusola, and S. Tang, "Perceptual color map in macaque visual area v4," The Journal of Neuroscience **34**, 202–217 (2014).
38. Y. Xiao, A. Casti, J. Xiao, and E. Kaplan, "Hue maps in primate striate cortex," NeuroImage **35**, 771–786 (2007).
39. B. Conway, S. Moeller, and D. Tsao, "Specialized color modules in macaque extrastriate cortex," Neuron **56**, 560–573 (2007).
40. A. W. Roe, L. Chelazzi, C. E. Connor, B. R. Conway, I. Fujita, J. L. Gallant, H. Lu, and W. Vanduffel, "Toward a unified theory of visual area V4," Neuron **74**, 12 – 29 (2012).