Spatio-chromatic Opponent Features

Ioannis Alexiou and Anil A. Bharath

BICV Group, Imperial College London, UK

Abstract. This work proposes colour opponent features that are based on low-level models of mammalian colour visual processing. A key step is the construction of opponent spatio-chromatic feature maps by filtering colour planes with Gaussians of unequal spreads. Weighted combination of these planes yields a spatial center-surround effect across chromatic channels. The resulting feature spaces – substantially different to CIELAB and other colour-opponent spaces obtained by colour-plane differencing – are further processed to assign local spatial orientations. The nature of the initial spatio-chromatic processing requires a customised approach to generating gradient-like fields, which is also described. The resulting direction-encoding responses are then pooled to form compact descriptors. The individual performance of the new descriptors was found to be substantially higher than those arising from spatial processing of standard opponent colour spaces, and these are the first chromatic descriptors that appear to achieve such performance levels individually. For all stages, parametrisations are suggested that allow successful optimisation using categorization performance as an objective. Classification benchmarks on Pascal VOC 2007 and Bird-200-2011 are presented to show the merits of these new features.

Keywords: Colour descriptors, image categorization, colour-opponency, biologically-inspired, pooling, Bird 200, Pascal VOC.

1 Introduction

In image classification, colour is often treated as an auxiliary feature that can be called on to boost classification rates. To be specific, a common approach to the incorporation of colour has been to fuse chromatic features with achromatic features such as SIFT [16]. Colour features on their own have, to date, not been shown *individually* to produce good classification performance. This might be attributed to the substantial information about image structure that is contained within achromatic gradients; it may also be that illumination variations and shadows can cause strong shifts in hue and saturation. Both of these factors may be important considerations. Clearly, with the exception of effects in perception that are induced by higher cognition, one would wish that, for two arbitrary image patches, descriptor similarities should follow perceptual similarity. It is, therefore, not unreasonable to explore how biological mechanisms of processing colour information might differ from existing techniques for generating colour descriptors. We first review the existing relevant literature on colour descriptors.

D. Fleet et al. (Eds.): ECCV 2014, Part V, LNCS 8693, pp. 81–95, 2014.

1.1 Related Work in Colour Description

Prior research on colour descriptors has included the effects of different colour representations on classification performance in standard datasets [18], [12]. This has led to four main variants of descriptor that are relevant to the work reported in this paper:

Colour SIFT. Previous work [18] has explored various colour-spaces in image classification, but used a common descriptor sampling approach: in essence, SIFT-based descriptors were applied on the channels of various colour spaces. The two feature types most relevant to the current paper may be described as “OpponentSIFT and “C-SIFT, and both are derived from opponent colour spaces. The difference between these two features is that C-SIFT makes use of C-invariants as suggested by [12] to provide confidence on the colour channels, whereas OpponentSIFT relies only on the raw colour-opponent channels.

HSV-SIFT. Another recent approach [2] computed SIFT descriptors over all three channels of the HSV color space, yielding one descriptor for each channel. HSV channels, however, produce a description which is not purely invariant to light intensity; different lighting conditions affect the colour encoding in the hue and saturation channels. This lighting sensitivity feeds into the descriptors produced by such an approach.

Hue-SIFT. The technique of [20] combines the achromatic SIFT descriptor with a hue histogram. The hue channel of HSV space is known to exhibit unstable behaviour for colour pixels that lie close to the grey axis of a bi-conical colour-space model. To address this, the implementation proposed in [20] uses the saturation values to weight the bins of the hue histogram. This weighting reduces the effect of low-confidence hue values, improving the reliability of the hue histogram over an unweighted version.

MS-SIFT. Multi-spectral SIFT [5] is an extension of SIFT into an opponent colour space. Four channels of information – the RGB and near infra-red channels – are decorrelated, producing a space that is closely related to the opponent-colour model. A SIFT-type feature is then constructed from the decorrelated data. The experiments discussed in [5] departed from the (currently) more widely used classification pipelines, in that sparse and scale-selective keypoints, rather than dense sampling, were used to assess the performance of the multi-spectral features.

SO-DO units. A biologically-inspired descriptor, proposed by Zhang *et al.* [22], imitates the colour processing thought to be found in the early stages of some mammalian visual systems. The SO-DO scheme employs colour opponent processing units that split the signed outputs of weakly oriented filters applied separately on colour channels; other authors [21,11] have used similar processing models. The single-opponency (SO) units were extended to double-opponency (DO) units by applying another set of oriented filters to each colour opponency channel obtained from the SO units. One critique of this system – at least from a biological perspective – is that *primate* vision applies opponent processing as early as the retina. Directionally-selective neuronal responses in primates only emerge (neglecting feedback) further in the feed-forward visual path, at the level of the primary visual cortex. Nevertheless, we consider this approach to be quite relevant, and its performance is discussed in Sec. 5.

Discriminative Colour Descriptors. A recent suggestion for incorporating colour information in classification is the approach of the so-called “discriminative color descriptors” [14], in which the spatial colour information is clustered according to criteria that minimize the mutual information of colour features obtained from the CIELAB colour space. The clusters are assigned to the original image using a bag-of-words approach to build up a classification pipeline.

2 Motivation

A digital image acquisition typically yields values in three channels of RGB colour space. Because of the broad wavelength sensitivity functions of pixel sensors and the spatial interpolation (for example, to compensate for Bayer pattern sampling) on many sensing devices, a change in the light falling at some point in the imaging plane will manifest itself on all three channels of nearby image pixels. Consequently, if gradient fields are computed for each of the three channels, the change introduced by illumination will be distributed across all bins of all gradient histograms. Decorrelating transformations, such as Principal Components Analysis (PCA) or Zero-Phase Components Analysis (ZCA) [6] can be used to remove this linear correlation, and it has been noted that the resulting transformed colour spaces appear quite similar to the so-called opponent colour space, which contain channels that explicitly encode *differences* in red-green and blue-yellow components.

Standard chromatic opponency is thought to encode spatial colour efficiently, and in simple colour grouping tasks leads to results that are more closely aligned with human

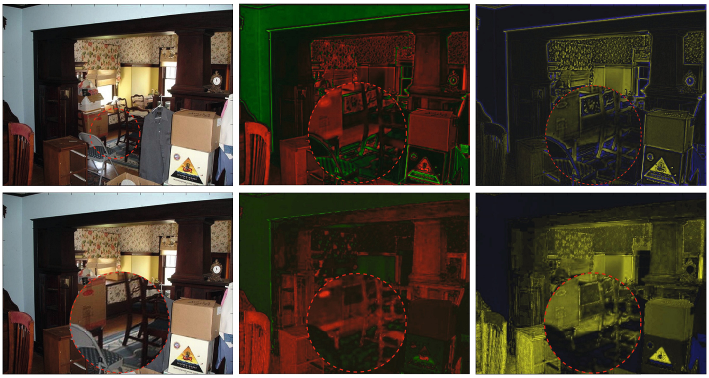


Fig. 1. A comparison illustrating the apparent loss of visual information from the raw $a*$ and $b*$ channels of CIELAB space relative to the equivalent Opponent Difference of Gaussians (OpDoG). The first image on the top row shows the original image and detail region. The top middle and top right images show the proposed opponent R-G and B-Y OpDoG channels. Both circled regions are at a zoomed scale so that detail can be seen. The leftmost image of the bottom row shows the detail of the original image at the zoomed scale; the middle bottom panel shows the R-G component of La^*b^* space (i.e. $a*$), and the right panel, the B-Y channel (b^*). Note (detail) the difference in responses near to edges.

perception than “raw” RGB space. A well-known opponent colour space is CIELAB, but there are related spaces, such as LUV. The CIELAB space is produced from the trichromatic colour space CIEXYZ. This trichromatic colour space is calculated by “projecting” raw RGB space to the perceptual colour triangle. The three main colour coordinates of the CIEXYZ descriptions are used in an antagonistic organisation in order to capture the two main colour opponencies of red-green, blue-yellow and a third channel which provides only luminance information.

Commonly used opponent colour spaces, such as CIELAB, bear some similarities to colour processing found in mammalian retinal physiology [13,7]. For example, an appropriate non-dynamic model to describe the mapping from photoreceptor activation through to the firing rates of retinal ganglion cells is a Difference-of-Gaussians applied to distinct colour channels. This model does indeed process colour channels in an antagonistic manner. However, the operation of smoothing with different spatial kernels and colour channel differencing is non-commutative. At this point, it is worth comparing biological spatial and wavelength colour opponency with CIELAB space to highlight the difference. Roughly, with CIELAB, or any other common opponent channel, pixel-wise differences are used to yield the opponent channels or planes. We have found that directional colour boundaries are visibly less distinctive in standard CIELAB space (see Fig. 1, bottom middle and right panels). On the other hand, by paying attention to the order of computation involving smoothing and channel differencing, we can preserve much chromatic boundary information (Fig. 1, detail, top middle and top right panels). By using a custom approach to building descriptors from a more biologically accurate colour opponent space, performance from chromatic channels is greatly improved. Descriptors can, of course still be constructed for CIELAB space, but the gradient information that would be captured is less likely to contain salient information. The evidence for this is also present in the literature in classification experiments reported elsewhere, see for example [18]. Often, low performance in colour space is addressed by the continued use of the achromatic channels in final classifiers.

Before detailing this “Opponent Difference of Gaussians” space, we will outline the main contributions of this paper. We first use a colour processing model with a center-surround (isotropic) structure in order to generate colour representation channels that capture spatio-chromatic opponencies. The approach taken to tune a series of parameters for opponency, gradient estimation and pooling is discussed in Sec. 4. Because of the introduction of this spatial filtering model which modifies the Fourier content of an image (see Sec. 3.1), a custom gradient-like field estimation method is required. A generalised form of transfer function, allowing more freedom than partial derivatives of a Gaussian, is used in place of gradient field calculations. New pooling patterns are then learned using an optimisation approach. We demonstrate the *individual* performance of these joint spatial and colour descriptors, then show that feature fusion adds further improvements to classification rates.

2.1 Modelling Biological Opponent Colour Channels

The peak firing rate of neurons with isotropic spatial receptive fields occurring early in the visual system can be roughly approximated in a variety of ways. Two common alternatives from computational neuroscience are *Difference*, (*not* “Derivative”) of Gaussians (DoG) and the *Laplacian* of Gaussian (LoG), also known in computer vision. In an algorithmic implementation of a luminance-only retinotopic model, either DoG or LoG functions could be spatially sampled to produce a convolution mask. If modelling luminance-only receptive field responses, the difference between these two options is not substantial, though one has a larger number of parameters to play with in the DoG model. When one is attempting to model biological colour-opponent channel processing, the difference between an LoG and a DoG is crucial, as we shall next see. For a *single* achromatic channel, the opponent model can be understood in terms of the center-surround spatial weighting, D, described in Eq. (1), using a two-dimensional coordinate vector $\mathbf{r}$, with respect to a centre $\mathbf{r}_0$:

$$D(\mathbf{r}|\mathbf{r}_0, \sigma_{ce}, \sigma_{su}) = A_{ce} \exp\left(- \frac{|\mathbf{r} - \mathbf{r}_0|^2}{2\sigma_{ce}^2} \right) - A_{su} \exp\left(- \frac{|\mathbf{r} - \mathbf{r}_0|^2}{2\sigma_{su}^2} \right) \qquad (1)$$

Because this Difference of Gaussians is isotropic, parameter subscripts "ce" and "su" refer to the centre and surrounding regions, respectively, around a central spatial location $\mathbf{r}_0$; $A_{..}$ refers to the amplitude scalings, $\sigma_{..}$ controlling spread, and $\sigma_{su} > \sigma_{ce}$.

In primates, biological colour-opponent processing has different colour channels contributing to the centre and surround regions of a single unit. This cannot be achieved by applying a single D (as in Eq. (1)) or LoG function to the *difference* of colour channels. This is because although convolution itself is commutative with addition (or subtraction), the fully distributive property of convolution suggests that, for most non-trivial functions, f, g, h_1, h_2 of one, two or a higher number of dimensions *in general*:

$$(f - g) * (h_1 \pm h_2) \neq f * h_1 \mp g * h_2 \qquad (2)$$

In the context of two-dimensional colour planes, f and g are arbitrary real-valued chromatically selective channels and h_1 and h_2 are pairs of spatial convolution masks; $*$ denotes two-dimensional convolution. For the inequality expressed in (2), equality can only be reached iff $h_1 = h_2$ or $f = g$, or any of the 4 operands on the Left Hand Side (LHS) is identically 0. The net effect is that the Right Hand Side (RHS) of (2), in which two different blurring functions h_1 and h_2 are applied to channels f and g respectively, captures different information to either sum or difference of spatial blurring functions applied to an opponent $(f - g)$ channel (LHS of (2)). This appears to be important.

A first step in our colour-opponent DoG descriptor is therefore obtained by convolving the raw colour channels with the two Gaussian functions, denoted by G_{ce} and G_{su}, for the central and surrounding colour planes, respectively:

$$OpDoG = W_c \bar{\times}_3 (I * G_{ce}) - W_s \bar{\times}_3 (I * G_{su}) \qquad (3)$$

The two Gaussian spatial kernels are applied through convolution across each of the three colour channels of an RGB image, denoted by I. The desired opponency channel is obtained by applying the two mixing vectors to the convolution outputs. In (3) we describe this with a tensor-vector product along mode 3, using the notation of Kolda [15], and treating both the $M \times N \times 3$ result of $I * G_{ce}$ and the $M \times N \times 3$ result of $I * G_{su}$ as order 3 tensors. The result of a tensor-vector product is one less than the order of the tensor, and so the terms to either side of the "-" sign are order-2 tensors, and may be treated as 2D scalar fields.

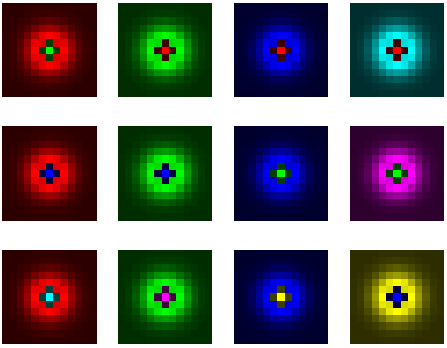


Fig. 2. This illustrates the opponent DoGs that have been used to encode chromatic contrast. The opponent colours were not restricted to Red-Green and Blue-Yellow, but allowed to vary in order to identify useful combinations. These are the actual kernel sizes used to generate the OpDoG results. Best viewed in colour; see text for details.

The terms W_c and W_s are both 3×1 mixing vectors containing the coefficients for the desired colour components of the opponency, and both take the form $[w_r, w_g, w_b]^{\mathrm{T}}$. For any pair of selected centre and surround mixing vectors, the output of Eq. (3) is therefore a single 2D array which incorporates a single opponency. There are, of course, many mixing combinations of these channels. A subset of all possible opponent DoGs was examined in this work, and these are presented in Fig. 2, which illustrates 12 opponent DoGs, organised into 4 columns. The first column has the red component as the surround characteristic. The central components on the first column use the remainder channels of green and blue, as well as their blend (cyan). In the second column, the green component is used in the surround, with the remaining channels (and their blends) forming the center components. The same approach was taken for the third and fourth columns.

3 From Opponency to Descriptors

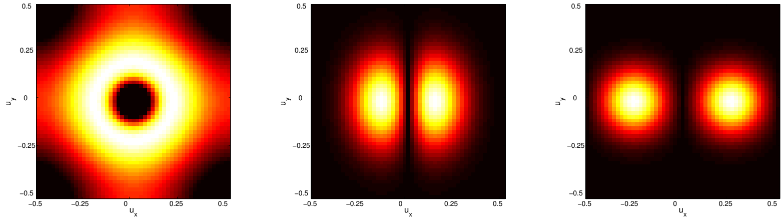
The feature maps from the opponent DoGs were further processed to capture directional information. In most single-channel descriptor constructions, derivative estimators are applied directly to the intensity channel to yield a gradient field. However, given that spatial filtering has already been applied to generate the OpDoG feature planes, the estimation of directional information from the "modified" image data requires an appropriate operator to be designed.

3.1 Directional Responses

Because of the approximate similarity of the OpDoG operator to an isotropic Laplacian, one might expect that its effect on a single opponent channel would be similar to a bandpass filter. However, it turns out that in order to achieve an overall (net) response that is closer to that of a Gaussian derivative, expressed in 2D Fourier space (u_x, u_y) as:

$$\hat{T}_{GDD} \propto u_x \; \exp\left[-\left(\frac{u_x^2}{2\sigma_{u_x}^2} + \frac{u_y^2}{2\sigma_{u_y}^2}\right)\right] \tag{4}$$

(see Fig. 3(b)), a spatial kernel is required such that the cascade of operators – OpDoG followed by some direction selective operators – will yield a field that encodes something similar to gradient direction in the relevant opponent channel. This has an almost direct biological analogy in the computational structure of higher mammals, in which afferent projections of neurons with isotropic receptive fields in the Lateral Geniculate Nucleus (LGN), a thalamic structure, are collected and weighted to yield direction-sensitive responses in visual cortex. Recognising this, we opted to take a more general approach to designing the subsequent gradient-field operators, doing so in the Fourier domain (see Fig. 3).



(a) DoG (b) GDD (c) TFSG

Fig. 3. This series of figures illustrates magnitude spectra in Fourier-domain (u_x, u_y). (a) The *approximate* effect of an OpDoG operator at a fixed slice in colour space; (b) a Gaussian Directional Derivative (GDD) imposes a spatial pattern that is clearly anisotropic. To build descriptors, (a) has to be processed to be direction selective, as in (b), and multiple directions must be synthesized. We found that the magnitude spectrum in (c) performs well, and is referred to as a *shifted gradient* (compare with (b)). Note that because opponency is applied across colour space, these illustrations are only approximate.

Fig. 3 (a) shows the effective range of frequencies that the OpDoG responses produce in Fourier space. Comparing Fig. 3 (a) with (b), it can be noted that a single-channel DoG Fourier response down-weights a large circular region in the middle (low frequencies); some of these frequency components are retained in a GDD operator. However, due to the combination of spectral bands in the OpDoG, it is unknown precisely which spatial frequency bands are modified, relative to a single-channel GDD. Thus, a flexible directional (i.e. tunable) transfer function (see Fig. 3 (c)) is proposed to produce a directional filter (Eq. (5)) when taken with the effect of the OpDoG. The new band pass selective filter (see Fig. 3 (c)) is applied on the output of OpDoG (Fig. 3 (a)), and falls inside the effective region of the OpDoG Fourier response; because it will have a frequency response quite different to a GDD, we refer to it as a *shifted* gradient operator.

We propose a Transfer Function of a generalised "Shifted Gradient", in the Fourier domain. The form of this function is:

$$\hat{T}_{TFSG} \propto \exp\left[-\left(\frac{|\log_\kappa u_x - \log_\kappa u_{x0}|^\gamma}{(\log_\kappa \sigma_{u_x})^\gamma} + \frac{|\log_\kappa u_y - \log_\kappa u_{y0}|^\gamma}{(\log_\kappa \sigma_{u_y})^\gamma}\right)\right] \tag{5}$$

The components of Fourier space of each OpDoG channel that correspond to antisymmetry in image space were extracted from the magnitude spectrum illustrated in Fig. 3 (c). The form of $\hat{T}_{TFSG}$ is appropriate to introduce orientation selectivity into Fourier space following the application of OpDoG filtering. It also allows parameter optimisation to be applied to improve classification performance. In Sec. 4, we will describe the range of the parameter values of Eq. (5) in order to produce directional responses. We will also describe the method used to optimise performance. Briefly, the final directional channel design was found by optimising the effect of the five parameters of the TFSG function on classification performance: the $\log_{\kappa}$ term in Eq. (5) modulates the radial skewness of the function; real parameter γ modulates the shape of the pass-band, changing its spatial kurtosis; parameter pair of (u_{x_0}, u_{y_0}) translates the band-pass regions and the parameter pair of $(\sigma_{x_y}, \sigma_{u_y})$ controls the width and the aspect ratio of the transfer function. A $\pi/2$ rotation of the pattern shown in Fig. 3(c) was also used to generate the second component of a directional field.

3.2 Pooling Patterns

Having generated directed responses for opponent colour space, we pursued an approach to build descriptors that would enable parameter tuning to be easily achieved. At the same time, we sought to keep the dimensionality of the resulting descriptors comparable to a SIFT-type approach. In addition to a new pooling approach, which we will now describe, we also applied SIFT grid and histogram-binning methods to produce descriptors from the post-filtered (i.e. TFSG-processed) OpDoG fields. Comparisons of performance between both of these pooling approaches are presented in Sec. 5.

The directional responses from the shifted gradient operators can be captured over discrete image space by applying a descriptor pooling scheme similar to that used in SIFT features [16] or a Gaussian arrangement of pooling sectors [19,4]. The SIFT descriptor uses histograms to describe the gradient patterns in a local region of image space; there is no sub-patch weighting when producing the descriptor entries (though there is for the overall patch orientation estimate). Yet, two other studies [19,4] have shown that application of local spatial weighting during pooling can improve the performance of a descriptor. Both of these studies applied Gaussian pooling functions. We wished to explore whether non-Gaussian patterns could be applied successfully.

Using the two-dimensional form of:

$$\hat{\Phi} = \exp\left[-\alpha \left|\log_b\left(\frac{x^2+y^2}{d_n^2}\right)\right|^p - \beta|\theta-\theta_m|^p\right] \tag{6}$$

we designed two-dimensional templates for use as pooling functions to encode the shifted spatial gradient outputs into the elements of colour patch descriptors. The terms on the RHS of Eq. (6) allow spatial kurtosis (p), skewness (log_b), spatial-scale (α and β) and translation (radial $\frac{(x^2+y^2)}{d_n^2}$ and angular $\theta - \theta_m$). The discrete indices (m, n) refer to pooling regions in angular ($m = 0, 1, ..., 7$) and radial ($n = 1, 2$) fashion. Fig. 4 illustrates the distributions after they have been tuned following the procedures to be described in Sec. 4.

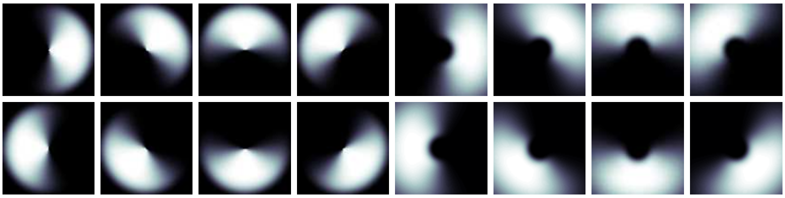


Fig. 4. These 16 spatial pooling patterns were learned from a set of training data. They are applied by Frobenius inner product to the outputs of eight directed channels of opponent colour, leading to comparable descriptor sizes to a "standard" SIFT descriptor. The borders of these patterns incorporate smooth weight decay. Novel factors of this design are the polar arrangement of these patterns, and their strong degree of spatial overlap. Each pooler produces 8 entries in the final descriptor for each OpDoG channel.

The poolers are applied over 16×16 patches spaced every 6 pixels, following TFSG filtering of each OpDoG channel. The operation between the pooling pattern and a channel patch of the same size as the pooler is a simple Frobenius inner product.

4 Optimisation

A classification pipeline was set up to use the Pascal VOC 2007 dataset in order to tune descriptor construction. A data selection process identified the minimum size of a subset of images from the training data that would lead to reliable performance increments through a classification module. This reduced the likelihood of overfitting, whilst removing the need to use all descriptors from all images during an intensive optimisation process. On the selected subset of images, the pooling arrangement was then tuned with the objective of maximizing classification performance, again on training data. The encoding of image to descriptor is described in the order of Sec. 2.1, 3 and 4, but the final design used the optimisation described in this Section.

The harvested features, each in the form of 128-dimensional vectors, are projected onto their 80 principal components to reduce dimensionality. The projected features are clustered by fitting 256 Gaussian models using a standard Gaussian Mixture Model. A diagonal covariance matrix structure was enforced. A spatial pyramid of two levels (0,1) and three horizontal stripes, similar to the approach of Van de Sande *et al.*[18], was used to define the descriptor-codebook relationships using Fisher vector encoding [17]. There are two separate learning stages, with the first stage learning the parameters discussed in Sec. 2.1 (by seeking in Eq. (3) the σ_{ce} and σ_{su} of G_{ce} and G_{su} within $(0.1, 2)$ with a stride of $\delta_{ce,su} = 0.25$) and Sec. 3 for the directional OpDoG channels. The second stage uses the learned parameters of the OpDoGs with subsequent gradient field approximation to learn the pooling patterns.

We used the mAP to tune all parameters using Powells multidimensional direction set method within the bounds (and a stride of $\delta_{(.)}$) outlined in Eq. (7) and Eq. (8):

$$\hat{T}_{TFSG} = \begin{cases} \kappa & : \{\kappa \in (0.1, 4), \delta_\kappa = 0.5\} \\ \gamma & : \{\gamma \in (0.1, 4), \delta_\gamma = 0.5\} \\ u_{x_0} & : \{u_{x_0} \in (0.1, 0.4), \delta_{u_{x_0}} = 0.05\} \\ u_{y_0} & : \{u_{y_0} \in (0.1, 0.4), \delta_{u_{y_0}} = 0.05\} \\ \sigma_{u_x} & : \{\sigma_{u_x} \in (0.1, 5.5), \delta_{\sigma_{u_x}} = 0.05\} \\ \sigma_{u_y} & : \{\sigma_{u_y} \in (0.1, 5.5), \delta_{\sigma_{u_y}} = 0.05\} \end{cases} \quad (7) \qquad \hat{\Phi} = \begin{cases} \alpha & : \{\alpha \in (0.01, 8), \delta_\alpha = 0.5\} \\ \beta & : \{\alpha \in (0.01, 8), \delta_\beta = 0.5\} \\ p & : \{p \in (0.5, 4.5), \delta_p = 0.5\} \\ d_m & : \{d_m \in (0.1, 0.9), \delta_{d_m} = 0.1\} \\ \theta_m & : \{\theta_m \in [0, 2\pi), \delta_{\theta_m} = \pi/4\} \end{cases} \quad (8)$$

The parameters that were learned for the TFSG improved performance over the GDD by 1.1% (mAP) in the Pascal VOC dataset. The new pooling patterns improved performance over SIFT-type pooling schemes by 2.3% (mAP) and 3.4% (mAP) for the Gaussian based configuration as ("Daisy") in [18,4] using 17 pooling regions. In Sec. 5, individual OpDoG channel performance is reported, as are comparisons of different pooling techniques and colour spaces.

5 Classification Benchmarks

Classification performance was assessed using a series of experiments designed to identify consistent causes (e.g. parameter settings) of improvement in categorization performance. This included parameters of the OpDoGs, the gradients, and the pooling patterns found in Sec. 4. We built a standard categorization pipeline for both the Bird-200 [3] and Pascal VOC 2007 [10] datasets. The parameters found through optimisation (see Sec. 4) using a VOC2007 training subset were evaluated on the VOC2007 test set and on the Bird-200 dataset without further optimisation. This shows that the parameter tuning does generalise, leading to satisfactory performance on a completely different dataset. Although we performed comparisons using standard smoothed gradient estimators ("SIFT" in the Tables), we also built an independent, parallel path of processing that allowed us to vary the mixing of the basic colour channels. In the performance results, these are referred to by the opponent components involved "-RG", "-RB", "-RC", "-GB", "-GM" and "-BY" (red-green, red-blue, red-cyan, green-blue, green-magenta, blue-yellow), all being produced from Eq. (3) and sampled by the pooling patterns presented in Fig. 4. The experiments were separated into single scale and multiscale versions, leading to different sets of results. This approach was necessary in order to tease out the nature of any performance differences, particularly as the complexity of the encoding increases by a factor of approximately 4 in moving from single-scale to multi-scale methods. To enable performance comparisons, the nearest relevant opponent filtering system – the SO method – is included in the Pascal evaluation. The performance in this test did not warrant further evaluation in the Bird-200 dataset.

Single/Multiscale. The *single scale* classification rates were obtained using the following setup. The low-level feature spaces resulting from the OpDoG channels were sampled using the pooling patterns shown in Fig. 4. The number of pooling regions (16) was selected to be as comparable to the commonly-used 128-element SIFT descriptor [16] as practical. For all descriptors, 16×16 patches were sampled and these were spaced (dense grid) every 3 pixels, as described in the relevant experimental section (referred to using the suffix "SNG"). The multiscale classification setup is performed by following a previously described [9] arrangement of 4 scales. For these experiments, the descriptor sampling density was fixed at 3 pixels. Similar to the experiments of Chatfield *et al.*

[9], the spatial pooling size of the SIFT descriptor for the multiscale case (referred to with suffix "MLT") was set to 4, 6, 8 and 10 pixels.

Classification Pipeline. For the experiments on both datasets, a Gaussian Mixture Model was employed to produce 256 components for use in Fisher-vector encoding [17]. A spatial pyramid of two levels (0,1) and 3 horizontal stripes [18] was also applied to allow comparison with other recent work. Finally, an SVM employing a Hellinger kernel was used for the Fisher vectors, to maintain consistency with other recent work [17]. In order to accommodate the different pyramidal levels in the classifier, the kernels generated from each level of the pyramids were averaged and fed into an SVM for each class. The testing protocol of Pascal VOC was used to report class-specific average precision. The authors of the Bird-200 dataset provide the splits for the training and testing without a specific classification measure. Thus, the mAP and per- class classification accuracy are reported and discussed.

5.1 Experiments on Pascal VOC 2007

The mean average precision (mAP) is provided for the Pascal VOC 2007 dataset [10]. In Fig. 5, results from six combinations of possible colour opponencies are presented in order to assess individual OpDoG channel performance. The performance of the OpDoGs is compared with a state-of-the-art approach which is based on the implementation described in [9] and is denoted as "SIFT-MLT". Actually, using this particular classification pipeline, two scale sampling approaches were taken and compared: a multiscale (SIFT-MLT) and a single scale (SIFT-SNG). The proposed colour features were used only in single scale and are directly comparable to SIFT-SNG. However, the new colour opponent channels are not used in a multiscale fashion because of the computational cost that would be incurred in Fisher vector encoding when combining multiscale features with multi colour-opponency.

Despite the lower performance of single-scale OpDoG-based descriptors relative to an achromatic SIFT-MLT, a more direct comparison is facilitated by using a single scale of the basic pipeline. For example, OpDoGs that include green chromatic channels perform better than SIFT-SNG, even though in some cases the relative improvements are marginal. To our knowledge, this is the first *comparable* colour feature with a performance that surpasses achromatic SIFT-SNG. Instead of extending single-scale OpDoG features to multiscale versions, we opted to use late fusion to seek performance boosts. The OpDoGs-FUSED feature is created by merging all OpDoG flavours with SIFT-MLT. The resulting performance is indicative of a complementary effect between chromatic and achromatic channels of processing. Although not shown in Fig. 5, it was found that merging SIFT-MLT with OpDoG-SNG-RG and OpDoG-SNG-GM yielded rates as high as 62% mAP – suggesting that green spatial/chromatic opponent channels significantly enhance performance.

Specific improvements may be assessed by noting how much each processing stage affects the performance; to keep the results succinct (see Sec. 4), we used average performance of all colour opponent flavours (lower half of Fig. 5). The lowest performing descriptor is C-SIFT [1], which is a standard dense-SIFT implementation applied on the colour channels modulated by the C-invariant as described in [1,18]. SplitGrad-SD

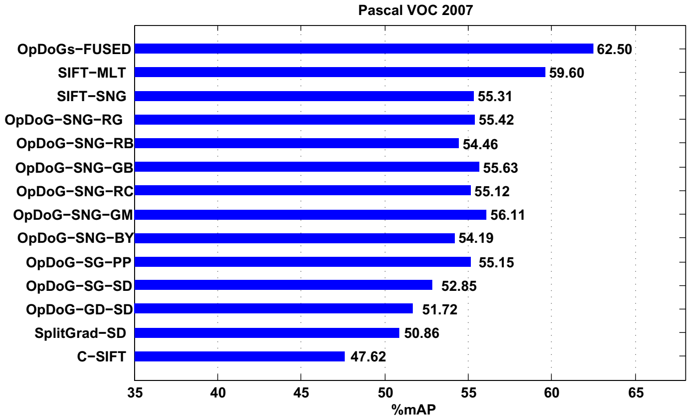


Fig. 5. The mAP is presented for each feature using the Pascal VOC 2007 dataset and protocol. This figure provides two sets of comparisons. The first presents the best performance of colour-opponent channels, including the use of feature fusion (from OpDoGs-FUSED to OpDoG-SNG-BY). The next set of results, starting from OpDoG-SG-PP and ending with C-SIFT [18] assesses changes in low-level processing modules, such as the gradient computation and the pooling schemes.

is quite similar to the SO units [22], but splits the positive and negative parts of gradient into two channels. Sampling is done with the SIFT approach i.e. a 16×16 grid and 128-bin histogram, "SD". The "OpDoG-GD-SD" feature represents the average of all colour opponent channels, Gaussian derivatives (referred to as "GD") and the SIFT descriptor (referred as "SD"). The feature "OpDoG-SG-SD" shows improved performance over "OpDoG-GD-SD" by replacing the gradient estimation with the shifted gradients (referred to as "SG"), described in Sec. 3. Finally, the "OpDoG-SG-PP" improves relative to "OpDoG-SG-SD" by replacing the SIFT descriptor with the pooling patterns (referred as "PP") from Fig. 4.

Comparing the rates of Fig. 5 with other recent results in the literature, we found the performance closest to ours was obtained by recent work of Khan *et al.* [14] (62% mAP); see also [18], which reports 56.6% mAP for C-SIFT. However, these approaches provide results of *fused* versions and not individual colour-channel performance; the proposed OpDoG features appear to stand out, significantly exceeding the individual channel performance of techniques such as the discriminative colour descriptors [14] (12% mAP).

5.2 Experiments on Bird-200-2011

The Bird-200 [3] dataset was selected for several reasons. First, it has a far larger number of sub-categories than Pascal VOC (200 species of bird - 11,788 images), and it is considered that colour and shape are equally important for fine-grained discrimination within this dataset. It is, thus, quite an appropriate challenge that is not overly dependent on either shape or colour features alone.

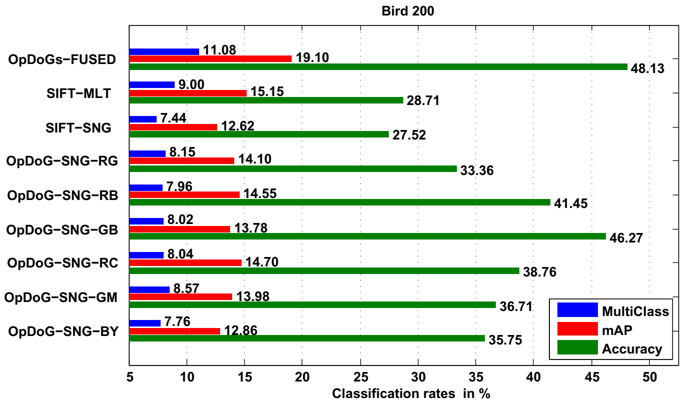


Fig. 6. Three different classification metrics were used to assess classification performance. The mAP rates are produced by using all of the training set, whilst the “Accuracy” and “Multiclass” measurements were obtained by randomly selecting 30 training examples for 10 iterations.

This dataset is also accompanied with suggested train-test splits, but the exact protocol for reporting the performance is optional. Thus, using the suggested data partitioning, three classification metrics are reported. In the red bars of Fig. 6, the mAP is reported as per the Pascal VOC protocol, using all of the training examples; classification accuracy (green bars) was calculated as per the Caltech 101 protocol (using 30 training examples), and Multiclass accuracy using 30 training examples, is displayed using blue bars. Baseline categorization performance is established with SIFT-MLT and SIFT-SNG to be able to identify relative improvements. In Fig. 6, all OpDoG features perform better than the SIFT-SNG. It is surprising that OpDoG-SNG-RB and OpDoG-SNG-RC perform closely to SIFT-MLT (comparing the mAPs), which is a feature with 4 times the computational effort of OpDoG and SIFT-SNG. Hence, this pair of features suggests that the OpDoG channels which capture red-opponent contrast are highly appropriate for this dataset. This claim is supported by noting that (not shown in

Table 1. Performance comparisons. Columns with (*) combine grey-scale and colour features, usually SIFT. Others are only colour. OpDoG is the fusion of all OpDoG channels, OpFused* combines these with SIFT-MLT. ALL uses all features in [18].

(a) VOC2007

Feature Type	OpFused*	OpDoG	ALL*[18]	C-SIFT*[18]	DCD*[14]	SODOSIFT[22]
mAP	62.5%	58.5%	60.5%	56.6%	12%-62%	46.5%

(b) Bird200

Feature Type	OpFused*	OpDoG	DCD*[14]	TriCos*[8]	C-SIFT[14]
Accuracy	48.1%	46.5%	26.7%	25.5%	21.1%

figures) the fused version of SIFT-MLT, OpDoG-SNG-RC and OpDoG-SNG-RB reached 18.4% mAP. The different OpDoG features are merged (as for Fig. 5) with SIFT-MLT so as to illustrate the complementary behaviour of these features when added to multiscale intensity descriptors (SIFT-MLT) which, on their own, only perform at an Accuracy of 28%.

The classification accuracy (green bars) is higher than other reported approaches such as 26.7% [14] and 25.5% [8]. One factor that is worth mentioning is that the updated dataset of "Bird-200-2011" was used in this work, instead of "Bird-200-2010", which is of half the size. During our experiments, it was found that a very small number of training images (e.g. 30) is insufficient to reveal discriminating behaviour in features, partly because of the variance in performance (±10% in Accuracy).

6 Conclusions

This work suggests a new method for generating colour descriptors. Spatio-chromatic channels are created by differences between chromatic channel pairs that have been smoothed with Gaussians of different widths. Second, directional responses are created, using a customised filter design in the discrete Fourier domain. Pooling functions are then applied to create descriptors. In order to tune the process behind the OpDoG channels, a learning approach was introduced. This learning approach was sufficiently general to allow performance tuning by altering the mixing and center-surround parameters of the OpDoGs, the subsequent gradient estimators, and the design of the spatial pooling patterns. To our knowledge, the resulting descriptors are the first chromatic-sensitive descriptors, i.e. capturing both chromatic and structural information, that yield high performance when used on their own. They were also amenable to clustering and dictionary generation, and when tested alongside multiscale chromatic features, appear to provide additional performance gains, showing that they contain complementary information. Also, it is worth repeating that some of the OpDoG features – for example, containing green channel opponency – appear to exceed the performance of standard achromatic SIFT in single-scale comparisons with a minimal computational effort (10 ms per image). A more general observation is that differences in feature performance cannot be reliably found using a CalTech-like testing protocol: larger amounts of data are needed in training, along with a performance measure such as mean-Average Precision (mAP).

References

1. Abdel-Hakim, A.E., Farag, A.A.: CSIFT: A SIFT descriptor with color invariant characteristics. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1978–1983 (2006)
2. Bosch, A., Zisserman, A., Muoz, X.: Scene classification using a hybrid generative/discriminative approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(4), 712–727 (2008)
3. Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., Belongie, S.: Visual recognition with humans in the loop. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 438–451. Springer, Heidelberg (2010)
4. Brown, M., Hua, G., Winder, S.: Discriminative learning of local image descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(1), 43–57 (2011)

5. Brown, M., Süsstrunk, S.: Multispectral SIFT for scene category recognition. In: Computer Vision and Pattern Recognition (CVPR), Colorado Springs, pp. 177–184 (June 2011)
6. Brown, M., Süsstrunk, S., Fua, P.: Spatio-chromatic decorrelation by shift invariant filtering. In: CVPR Workshop on Biologically Consistent Vision (WBCV 2011), Colorado Springs, pp. 9–16 (June 2011)
7. Buzás, P., Kóbor, P., Petykó, Z., Telkes, I., Martin, P.R., Lénárd, L.: Receptive field properties of color opponent neurons in the cat lateral geniculate nucleus. The Journal of Neuroscience 33(4), 1451–1461 (2013)
8. Chai, Y., Rahtu, E., Lempitsky, V., Van Gool, L., Zisserman, A.: TriCoS: A tri-level class-discriminative co-segmentation method for image classification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 794–807. Springer, Heidelberg (2012)
9. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: Proceedings of the British Machine Vision Conference, BMVC (2011)
10. Everingham, M., Gool, L., Williams, C.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) challenge. International Journal of Computer Vision 88(2), 303–338 (2010)
11. Gao, S., Yang, K., Li, C., Li, Y.: A color constancy model with double-opponency mechanisms. In: IEEE International Conference on Computer Vision (ICCV), pp. 929–936. IEEE (2013)
12. Geusebroek, J.-M., Van den Boomgaard, R., Smeulders, A.W.M., Geerts, H.: Color invariance. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(12), 1338–1350 (2001)
13. Johnson, E.N., Hawken, M.J., Shapley, R.: The orientation selectivity of color-responsive neurons in macaque V1. The Journal of Neuroscience 28(32), 8096–8106 (2008), doi:10.1523/JNEUROSCI.1404-08.2008
14. Khan, R., Van de Weijer, J., Khan, F.S., Muselet, D., Ducottet, C., Barat, C.: Discriminative color descriptors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2866–2873. IEEE (2013)
15. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Review 51(3), 455–500 (2009)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
17. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the Fisher vector: Theory and practice. International Journal of Computer Vision 105(3), 222–245 (2013)
18. Van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1582–1596 (2010)
19. Simonyan, K., Vedaldi, A., Zisserman, A.: Descriptor learning using convex optimisation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 243–256. Springer, Heidelberg (2012), doi:10.1007/978-3-642-33718-5-18
20. van de Weijer, J., Gevers, T., Bagdanov, A.D.: Boosting color saliency in image feature detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(1), 150–156 (2006)
21. Yang, K., Gao, S., Li, C., Li, Y.: Efficient color boundary detection with color-opponent mechanisms. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2013)
22. Zhang, J., Barhomi, Y., Serre, T.: A new biologically inspired color image descriptor. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 312–324. Springer, Heidelberg (2012), doi:10.1007/978-3-642-33715-4-23