Discriminative Color Descriptors

Rahat Khan[1], Joost Van de Weijer[2], Fahad Shahbaz Khan[3], Damien Muselet[1], Christophe Ducottet[1] and Cecile Barat[1]
[1]Université de Lyon, F-42023, Saint-Étienne, France
CNRS, UMR 5516, Laboratoire Hubert Curien, F-42000, Saint-Étienne, France
Université de Saint-Étienne, Jean-Monnet, F-42000, Saint-Étienne, France
[2]Computer Vision Center, Barcelona
[3]Computer Vision Laboratory, Linköping University, Sweden
rahat.khan@univ-st-etienne.fr

Abstract

Color description is a challenging task because of large variations in RGB values which occur due to scene accidental events, such as shadows, shading, specularities, illuminant color changes, and changes in viewing geometry. Traditionally, this challenge has been addressed by capturing the variations in physics-based models, and deriving invariants for the undesired variations. The drawback of this approach is that sets of distinguishable colors in the original color space are mapped to the same value in the photometric invariant space. This results in a drop of discriminative power of the color description.

In this paper we take an information theoretic approach to color description. We cluster color values together based on their discriminative power in a classification problem. The clustering has the explicit objective to minimize the drop of mutual information of the final representation. We show that such a color description automatically learns a certain degree of photometric invariance. We also show that a universal color representation, which is based on other data sets than the one at hand, can obtain competing performance. Experiments show that the proposed descriptor outperforms existing photometric invariants. Furthermore, we show that combined with shape description these color descriptors obtain excellent results on four challenging datasets, namely, PASCAL VOC 2007, Flowers-102, Stanford dogs-120 and Birds-200.

1. Introduction

Local-feature based image representations have been successful in many computer vision applications, such as object recognition, image matching, and image retrieval. In many of these applications the local features are discretized into a visual vocabulary, which allows to represent images as histograms over visual words. In such representations, color next to shape, was found to be an important cue [16, 22]. In this paper we propose a new method to learn discriminative color descriptors.

Color description is difficult due to the many scene accidental events which influence its measurement. These events include shadows, illuminant changes, variations in scene geometry and viewpoint, and acquisition device specifications. This has sparked an extensive literature on photometric invariance which aims to describe color invariants with respect to some of these variations [12]. Based on reflection models [20] or assumptions on the illumination [8] invariance with respect to shadow, shading, specularities and illuminant color can be obtained. However, photometric invariance is gained at the cost of discriminative power. Therefore, in designing color representations it is important to weight the gains of photometric invariance against the loss in discriminative power.

An alternative way of describing color is by means of color names. Color names are linguistic labels humans use to communicate the colors in the world. Examples of color names are for example 'red', 'black' and 'turquoise'. Van de Weijer et al. [23] have proposed a method to automatically learn the eleven basic color names of the English language from Google images. The result of this learning is a partition of the color space into eleven regions. Then, an eleven dimensions local color descriptor can be deduced simply by counting the occurrence of each color name over a local neighborhood. Analyzing the clusters of RGB values which are appointed to a color name, let us consider 'red' for example, we note that these clusters possess a certain amount of photometric invariance. Multiple shades of red are all mapped to the same color name 'red'. However, when moving towards darker 'reds', at a certain point the values will

be mapped to the color name 'black' instead, and the photometric invariance breaks down. Recently, color names were found to compare favorably against photometric invariant descriptions on several computer vision applications, such as image classification [16] and object detection [14]. These results show that focus on photometric invariance which is at the basis of many color descriptors might not be optimal. They further suggest that discarding discriminative power of the color representation will deteriorate final results.

We propose to learn color descriptors which have optimal discriminative power for a specific classification problem. The problem of learning a color descriptor is equal to finding a partition of the color space. Our approach relies on the Divisive Information-Theoretic Clustering (DITC) algorithm proposed by Dhillon *et al*. [6] to learn this partition. We adapt this algorithm to ensure that the final clusters are smooth and connected. Considering all the values in the L*a*b* cube, we aim to join values in this L*a*b* cube driven by the discriminative power of the final representation, the latter being measured using information theory. We distinguish two variations. Firstly, the specific color descriptor which is optimized for a single data set. Secondly, a universal color descriptor which is trained on multiple data sets, thereby representing a wide range of real-world data sets. The advantage of universality is that users can run the learned mapping for an unknown data set without the effort of learning a data set specific color representation. In experimental results we will show that these discriminative color descriptors outperform purely photometric color descriptors, and that combined with shape description they can obtain state of the art results on several data sets.

2. Photometric Invariance versus Discriminative Power

Color feature design has been mainly motivated from photometric invariance perspective [10, 11]. It is based on the observation that colors in the world are dependent on scene incidental events such scene geometry, varying illumination, shadows, and specularities. To obtain invariance with respect to these effects, photometric invariant features can be derived. Often the dichromatic reflection model [20] is used to derive these invariances:

$$\vec{f} = m_b \vec{c_b} + m_i \vec{c_i} \tag{1}$$

where $\vec{f} = (R, G, B)$ is the pixel value. The color of the body reflectance is given by $\vec{c_b}$ and the surface reflectance by $\vec{c_i}$, m_b and m_i are scalars representing the corresponding magnitudes of the body and surface reflectance. For objects with matte reflectance, for which $m_i = 0$, it can for example be shown that $\vec{c} = \vec{f} \Big/ \left\| \vec{f} \right\|$ is invariant for m_b and hence for shadow-shading variation. Similarly, it can be shown that for specular surfaces the hue, another popular color descriptor, is invariant for specularities [12].

But one could wonder what the cost of photometric invariance is. Mapping multiple RGB values to the same photometric invariance will potentially lead to a drop in discriminative power. This aspect of photometric invariance has received relatively little attention. Stability and noise sensitivity where measured by Stokman et al. [13]. Geusebroek et al. [11] showed that with increasing invariances fewer Munsell patches could be distinguished. Here we will analyze the drop in discriminative power in a more principled way by means of information theory.

We discretize our initial color space into m color words $W = \{w_1, ..., w_m\}$. In our case m is equal to $m = 10 \times 20 \times 20 = 4000$ of equally spaced grid points in the L*a*b* cube. Consider we have a data set with l classes $C = \{c_1, ..., c_l\}$. These classes are represented by histograms over the color words. The discriminative power of the color words W on the problem of distinguishing the classes C can be computed by the mutual information:

$$I(C, W) = \sum_i \sum_t p(c_i, w_t) \log \frac{p(c_i, w_t)}{p(c_i)\, p(w_t)} \tag{2}$$

where the joint distribution $p(c_i, w_t)$ and the priors $p(c_i)$ and $p(w_t)$ can be measured empirically from the dataset.

The mutual information measures the information that the words W contain about the classes C. Now consider we join the words W into k clusters $W^C = \{W_1, ..., W_k\}$ which are invariant with respect to some physical variation. Each cluster W_i represents a set of words. Then Dhillon et al.[6] proved that the drop of mutual information caused by clustering a word w_t to cluster W_j (in our case based on photometric invariance) is equal to:

$$\Delta i = \pi_t KL\left(p(C|w_t), p(C|W_j)\right) \tag{3}$$

where the Kullback-Leibler (KL) divergence is given by

$$KL(p_1, p_2) = \sum_{x \in X} p_1(x) \log \frac{p_1(x)}{p_2(x)} \tag{4}$$

and $\pi_t = p(w_t)$ is the word prior.

The above Eq. 3 provides a way to assess for each color value the drop in discriminative power Δi which is caused by imposing photometric invariance. In Figure 1 we plot the drop in mutual information which occurs when we look at a photometric invariant representation with respect to luminance. This is simply obtained by defining clusters as the set of bins of equal (a, b) values, computing the $p(C|W_j)$ of each cluster, and computing Δi with Eq. 3. We plot the drop in mutual information as a function of lightness L and saturation $sat = \sqrt{(a^2 + b^2)}$. The plot is based on the Flower data set [19] but similar results were observed for other data

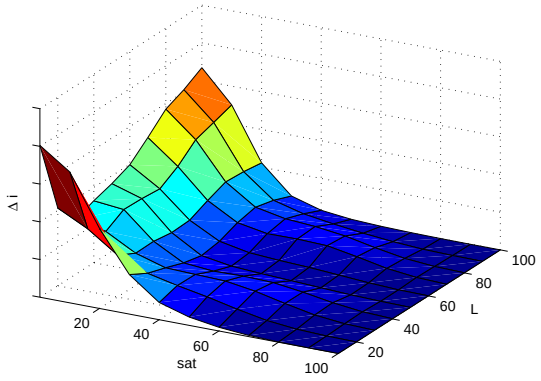


Figure 1. Graph showing the drop in mutual information for the flower data set caused by grouping bins with equal chromatic values (a and b). From the graph it can be seen that the drop of mutual information is largest for low saturated points, especially with low and high lightness (L).

sets. The plot tells a clear story: the largest loss of discriminative power is occurring for achromatic (or low saturated) colors as is clear from the ridge at $sat = 0$. Even though these achromatic colors cannot be distinguished from a photometric invariance point of view (since they can be generated from each other by viewpoint or shadow variations), this analysis shows that they contain discriminative power.

This leads us to investigate an alternative approach to color feature computations based on discriminative power. In the next section we outline our approach of discriminative color feature computation, which clusters color values together based on discriminative power on a training data set. The expectance is that discriminative clustering will automatically lead to a certain amount of photometric invariance: clustering values of similar hue together. However, in these regions — especially around the achromatic axis — we expect additional clusters to arise, to reduce the drop in discriminative power caused by the clustering.

3. Discriminative Color Representations

In this section we discuss our discriminative approach to color representations learning. We first explain divisive information-theoretic feature clustering (DITC) proposed by Dhillon et al.[6]. Next, we adapt the algorithm to find connected clusters in L*a*b* space.

3.1. DITC algorithm

The DITC algorithm provides an algorithm to cluster features into a smaller set of clusters, where each cluster contains a number of features from the original set. The clustering is performed in such a way as to minimize the decrease of mutual information (see Eq. 2) of the new more compact representation. The total drop of mutual information caused by clustering the words, using Eq. 3, is equal to

$$\Delta I = \sum_j \sum_{w_t \in W_j} \pi_t KL\left(p\left(C|w_t\right), p\left(C|W_j\right)\right). \quad (5)$$

Hence the clusters W which we seek are those which minimize the KL divergence between all words and their assigned cluster (weighted by the word prior). In our case the words represent L*a*b* bins of the color histogram. This color space is used because of its perceptual uniformity. Minimizing Eq. 5 is equal to joining bins from the L*a*b* histogram in such a way as to minimize the ΔI. L*a*b* bins which have similar $p\left(C|w_t\right)$ are joined together.

An EM like algorithm is used to optimize the objective function 5. The algorithm alternates between two steps.

1. Compute the cluster means with

$$p\left(C|W_j\right) = \sum_{w_t \in W_j} \frac{\pi_t}{\sum_{w_t \in W} \pi_t} p\left(C|w_t\right). \quad (6)$$

2. Assign each word to nearest cluster according to

$$\mathbf{w}_t^* = \arg\min_j KL\left(p\left(C|w_t\right), p\left(C|W_j\right)\right). \quad (7)$$

The new cluster index for word w_t is given by $\mathbf{w}_t^*$.

The algorithm is repeated until convergence. For more details we refer to [6].

The DITC algorithm has been studied in the context of joining color and shape features into so-called Portmanteau Vocabularies by Khan et al. [15]. In this paper, we use the DITC algorithm for a different purpose, namely to automatically learn discriminative color features. In addition, we propose two adaptations to the DITC algorithm.

3.2. Compact Color Representations

The original DITC clustering algorithm does not take into account the position in the L*a*b* space of the words. As a consequence, the algorithm can join non-connected bins. It is known that photometric variations result in connected trajectories [24]. Therefore when learning photometric invariants we expect them to be connected. In addition, connectivity has several conceptual advantages: it allows for comparison to photometric invariance, comparison with color names (CN), semantic interpretation (human color names are connected in Lab space), and comparison with human perception (e.g. MacAdam Ellipses). Therefore we propose to adapt the DITC algorithm to ensure that the clusters are connected in L*a*b* space. As a second adaptation we enforce smoothness of the clusters which prevents them from overfitting to the data. Both objectives can be translated into an additional energy term which can be added to the objective function of Eq. 5.

Let $\mathbf{w}_t$ be the cluster number assigned to word w_t, and $W_{\mathbf{w}_t}$ is the cluster to which w_t is assigned, then the cost of choosing a certain cluster assignment according to Eq.5 is equal to

$$\psi_t^I(\mathbf{w}_t) = \pi_t KL\left(p\left(C|w_t\right), p\left(C|W_{\mathbf{w}_t}\right)\right). \quad (8)$$

In this standard objective function, the relation of the words is not taken into account, and the final clusters W^C can — and most likely will — contain words which are not connected in color space. We enforce connectivity by introducing a cost for not being connected to the principle component of the cluster. The principle component $\mathcal{P}_j$ of a cluster W_j is defined as the connected component with the highest prior mass (the component for which the sum of the priors of its words is largest). Words which are not connected to the principle component of the cluster will have an additional cost for taking on this cluster assignment. We identify words connected to the principle component by $\mathcal{P}'_j$ and they are computed with a morphological dilation with a 26-connected structuring element b:

$$\mathcal{P}'_j = \mathcal{P}_j \oplus b. \quad (9)$$

This type of dilation is justified because we use equi-quantized bins on a uniform L*a*b* color space. After this dilation $\mathcal{P}'_j$ contains all words connected to the principle component of cluster j. We add a penalty term to all the color bins which are not part of $\mathcal{P}'_j$ according to

$$\psi_t^C(\mathbf{w}_t) = \alpha_C \cdot (1 - f^t(\mathbf{w}_t)) \quad (10)$$
$$\text{Where} \quad f^t(\mathbf{w}_t) = 1 \quad \text{if} \quad w_t \subset \mathcal{P}'_{\mathbf{w}_t}$$

With a sufficiently high choice of the constant α_C , this energy will eliminate non-connected assignments, and result in a final clustering of the features into connected clusters.

To enforce our second objective of smoothness of the color representation we introduce a pairwise cost according to

$$\psi(\mathbf{w}_s, \mathbf{w}_t) = \begin{cases} 0 & \text{if } \mathbf{w}_s = \mathbf{w}_t \\ \alpha_D & \text{otherwise} \end{cases} \quad (11)$$

Now consider a certain labeling for all words $\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_m\}$ then the cost of this labeling can be written to be

$$E(\mathbf{w}) = \sum_t \left(\psi_t^I(\mathbf{w}_t) + \psi_t^C(\mathbf{w}_t)\right) + \sum_{(s,t)\in\varepsilon} \psi(\mathbf{w}_s, \mathbf{w}_t) \quad (12)$$

where ε is the set of all connected words s and t.

The two step algorithm has to be slightly adapted to minimize this objective function. Step one remains unchanged and computes the cluster means. In step two we aim to find $\mathbf{w}^*$ which minimizes Eq. 12. This can be done with a graph cut algorithm where the nodes are the words (or bins of L*a*b* histogram) and the vertices connect neighboring nodes. After the optimal assignment $\mathbf{w}^*$ is found, the algorithm returns to step one until convergence.

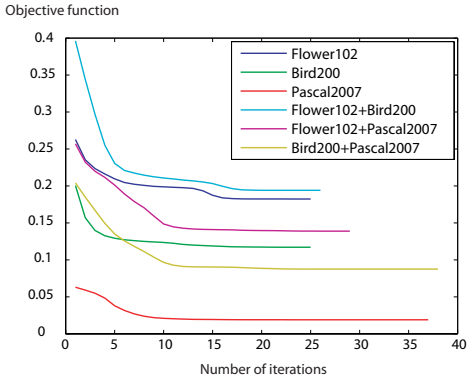


Figure 2. Evolution of the objective functions for some image sets until convergence.

3.3. Convergence

Our optimization of the objective function of Eq.12 is obtained by iteratively applying the two steps above. However, when we dilate all clusters (to define the connected bins), it could theoretically happen, that for some bins which change label, the bin to which they were connected also changes label. This could lead to unconnected components, and would activate the cost defined in Eq.10, and lead to an increasing objective function. This could be addressed by changing labels one bin at a time, but this would be computationally very costly. Practically, we run the iterations until no change in the labeling occurs. For the three datasets (and their three combinations) used in this paper, we verified that the final color descriptors were connected. Figure 2 shows the evolution of the objective function for the six runs until convergence.

3.4. Photometric Invariance of Learned Clusters

Instead of imposing photometric invariance, as is generally done, we follow an information theoretic approach which maximizes the discriminative power of the final representation. The underlying idea being that clustering color bins based on their discriminative power would automatically learn a certain degree of photometric invariance. Here, we verify that this has happened by analyzing the cluster assignments for two images.

We learn a 11-dimensional discriminative color descriptor for the Flower data set. Next, we apply the descriptor on two images of the data set. The results are depicted in Figure 3. Here, we replace the color of each pixel by the average color of all the pixels assigned to the same cluster. We can see that clusters are constructed so that they allow to discriminate flowers from background and leaves while providing some robustness across some photometric varia-



Figure 3. Examples of cluster assignment on two images from the Flower dataset.

tions. For example, note that the pixels under the shadows caused by the wrinkles on the yellow petals are assigned to the same cluster and the stamen part of the red flower is mapped to one cluster in spite of the photometric variations in the pixels. Also, the dark pixels that introduce most noises into photometric invariance representation are assigned to a separate cluster. The photometric invariance can also be observed from the bottom row of Fig. 5 where we see that pixels with similar hue but varying intensity are grouped together.

4. Universal Color Descriptors

In a seminal work named 'Basic color terms: their universality and evolution' the linguists Berlin and Kay [2] show the universality of the human basic color names. With universality they refer to the fact that the basic color names which are used in different cultures have a similar partition of the color space: the Arab azraq refers to a similar set of colors as the English blue. In the context of descriptors, we will use the term universality to refer to descriptors which are not specific to a single data set. Universality is one of the more attractive properties of the computational color names [23][1]. As a consequence of universality, users are not required to learn a new color representation for ever new dataset and can just apply the universal color representation to their problem.

In the previous section, we showed how to learn discriminative color features. Applying the above algorithm to a specific data set results in a color representation which is data set specific in the sense that it is optimized to discriminate between the classes of that data set. The same setup can be used to learn universal color vocabulary by joining several training sets together to represent the real-world. We learn such a description combining the training sets of Flower102, Bird200 and PASCAL 2007 data sets. An advantage over the existing computational color names [23] is that we are not limited to eleven color names and can freely



Figure 4. Example images from the four data sets used in this paper. From top to bottom: PASCAL 2007, Birds-200, Flowers-102, Dogs-120.

choose the desired dimensionality. We make the universal color descriptors available for the settings with 11, 25, and 50 clusters [1].

In the experiments we will investigate universal color descriptors, and compare them to specific color descriptors. We will do so by training the universal color descriptor from other data sets than the one currently considered. Universality is expected to result in a drop of performance since the descriptor cannot adapt to the specificity of the dataset. However, if the drop is small the advantages of a universal representation can outweigh the drop in performance.

5. Experimental Results

In the next few subsections, we discuss experimental details and results. At first, we briefly discuss the experimental setup and the details of discriminative descriptor learning. Then, we compare our proposed color descriptor with several photometric color descriptors on three image datasets. Next, we focus on the universality aspect of our descriptor and compare universality with specificity. In our final experiments, we combine our descriptor with shape description and compare results to the state of the art.

5.1. Experimental Setup

In this section, we briefly discuss the experimental setup used for sections 5.2 and 5.3. For these two sections we

[1]Example software and universal descriptor can be download from http://cat.uab.es/ joost/software.html

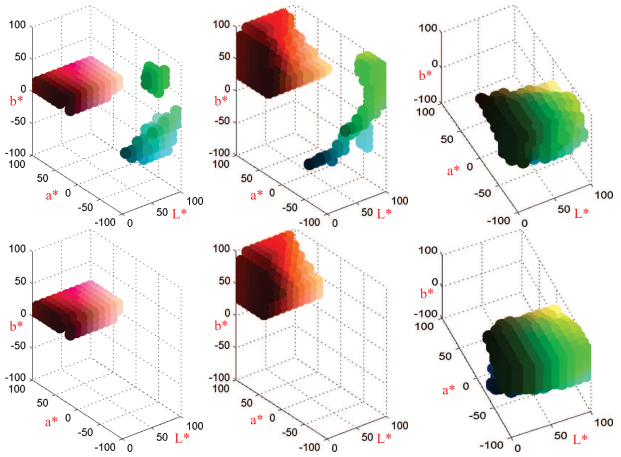


Figure 5. The clusters of the first and second row are computed from the Flower102 training set, by the original DITC algorithm and the proposed method respectively. Note the compactness and smoothness of the color clusters computed by the proposed method.

use a comparatively simpler framework to reduce the computational time, as our goal is to assess relative performance. For both sections, we choose three challenging image datasets, namely, Flower102 [19], Birds200 [25] and PASCAL 2007 (see Figure 4). For Flowers and Birds, the colors over the object classes are relatively constant. However for PASCAL 2007, colors are likely to change significantly in between samples of the same class (consider e.g. cars). In these experiments, we use a regular dense grid (16×16) with 50% overlap to extract patches from the images. After description of the patches, we employ a K-means on a random subset of features from the training set to build the visual vocabulary. We use SVM with an intersection kernel to obtain the classification score. The training and test set selection is consistent with the corresponding cited articles for each dataset. For section 5.4, we use a different experimental setup which is discussed in the beginning of that section.

For descriptor learning, for each dataset we convert all the training images from sRGB to L*a*b and construct a 3D histogram quantizing the L*a*b space by $10 \times 20 \times 20$, then we convolve these 3D histograms using a gaussian filter ($sigma = 1$). They are then used as 4000 dimensional feature vectors. We adapt the DITC implementation from [7] and use the Graph Cut implementation from [9]. As discussed in section 3.2, there are two parameters in our descriptor learning, namely, the dilation and smoothness cost parameters. The dilation cost parameter should be ideally equal to infinity, so we use a large enough value for that. Empirically we found that a smoothing cost parameter $\alpha_D = 10^{-8}$ obtained satisfying results on all data sets, and kept it constant.

We compare the clusters computed with standard DITC to the clusters computed with our algorithm which enforces connectivity and smoothness of the clusters. In Figure 5, we can clearly see that our method produces connected and smooth clusters. Note that, non-connected green parts from the first two clusters are associated to the green cluster when our method is employed. DITC only concerns discriminative clustering and does not ensure connected clusters which is undesirable from a colorimetric point of view.

Method	Flower102	Bird200	Pascal2007
rg	38.6%	4.3%	10.6%
HH	32.8%	3.5%	10.1%
CN	40.2%	7.7%	11.6%
DD(11)	43.7%	8.0%	12.2%
DD(25)	47.0%	8.7%	12.6%

Table 1. Comparison with photometric invariants.

5.2. Discriminative Color Descriptors

The aim of this paper is to arrive at a better color descriptors for object recognition directly on the discriminative power of the final representations. We start by comparing our discriminative descriptor(DD) to other pure color descriptors and the color name descriptor [23]. Note that in several comparisons color names were found to outperform various other pure color descriptors [16][14].

We consider two well known photometric invariants: normalized RGB (rg histogram) and a hue histogram (HH) [2] and the Color Names(CN) [23] [3]. We compare them against our descriptor with two settings, namely 11 and 25 clusters. Table 1 contains the experimental results. For each dataset we show the classification accuracy (or mean average precision for PASCAL 2007). For the case of 11 dimensions (equal to the CN descriptor) our descriptor obtains improved results on Flower and Bird, but slightly lower results than color names on PASCAL 2007. We can see from the table that our descriptor with 25 dimensions outperforms all the other descriptors used in the experiment. Note, that it is unclear how to increase the dimensionality of the color name descriptor above the eleven basic color names.

5.3. Universality versus Specificity

We discussed universality color descriptors because of their ease of use in section 4. In general, there is a growing interest in across-dataset generalization of methods in the community [21]. Here we use again the three datasets. We follow a leave-one-out approach, where we learn our descriptor on two datasets and test on the other. We also do dataset specific experiments, where we learn on one dataset

[2] Implementation provided by K. van der Sande at http://koen.me/research/colordescriptors

[3] As a sanity check we performed a k-means based LAB descriptor. Results were found to be inferior

and test on the same. In each case, we learn 3 different cluster groups i,e $k = [11, 25, 50]$ using our proposed method. We follow similar setup as section in 5.2 to represent images as bag-of-words.

It is evident from figure 6 that for larger k, the difference between universality and specificity becomes smaller. Also note that, the best results obtained using our universal descriptor, although not better than the specific ones, outperform other state-of-the art color descriptors used in experiments of section 5.2. In conclusion, for larger dimensions the drop of performance due to universality is relatively small, and users could prefer using it, rather than having to train a new dataset specific descriptor.

5.4. Discriminative Descriptors vs State-of-the-Art

We compare our approach with the state-of-the-art approaches in the literature. The experiments are performed on Birds-200, Flowers-102 and PASCAL 2007. Additionally, we also show the applicability of our approach on the challenging Stanford-Dogs 120 dataset. For our final experiments, we followed the standard bag-of-words pipeline. For feature detection, we use a combination of multi-scale grid with interest point detectors. For shape we use the SIFT descriptor. A visual vocabulary of 4000 is constructed for shape representation. For color, we use a visual vocabulary of 500 words. The vocabularies are constructed using standard K-means and the histograms are constructed using hard assignment. To represent an image we use the spatial pyramid representation as in [18]. For classification, we use the non-linear SVM using the χ^2 kernel [26]. We also compare our approach with the ColorSIFT descriptors [22] on the PASCAL VOC 2007, Birds-200 and Flowers-102 datasets. We use CSIFT descriptor for the PASCAL VOC 2007 dataset and OpponentSIFT for the other two datasets. A visual vocabulary of 4500 is constructed for ColorSIFT descriptors and an image is represented by spatial pyramids.The results are summarized in Table 2.

On the Birds-200 dataset shape alone provides a classification performance of 15.3. Our final result is a combination of late fusion between discriminative color and shape, shape alone and color alone. On this dataset our discriminative approach achieves the best classification score of 26.7 outperforming the colorSIFT [22] based on the same detected features. The universal color names result in a slight drop in performance. The other approaches in Table 2 also use a combination of color and shape. The portmanteau approach employ both color and shape to learn a compact color-shape vocabulary. The tricos approach [4] uses segmentation technique whereas for image representation shape and color with fisher vectors are employed.

On the Flowers-102 dataset, a mean accuracy of 69.0 is obtained. The incorporation of proposed color approach together with shape leads to 81.3. The universal color descriptor learned on the PASCAL 2007 and Birds-200 dataset results in a slight drop in performance. On this dataset again, our approach provides a comparable results to the state-of-the-art approaches in literature [3, 4, 19, 15]. On the PASCAL 2007 dataset, our framework with shape alone provides a meanAP of 59.9. Adding color with shape increases the meanAP to 62.0. The universal color descriptor results in slight deterioration in performance with a meanAP of 61.7. Again on this dataset, our final results are comparable to state-of-the-art results in literature [16, 22, 5, 27]. The method of [16] uses color attention approach to combine with color and shape with a meanAP of 58.0. The best reported results of 64.4 [27] is obtained using a different coding technique. Note that in this paper we use the standard vector quantization with hard assignment. However, our color descriptor can be used in any encoding framework together with SIFT.

Method	Birds-200	Flowers-102	Pascal 2007	Dogs-120
Tricos [4]	25.5	85.2	-	26.9
Bicos [3]	23.7	**85.5**	-	25.7
portmanteau [15]	22.4	73.3	-	-
Color Attention [16]	-	-	58.0	-
MKL [19]	-	72.8	-	-
LLC [17]	-	-	-	14.5
Fisher [5]	-	-	61.7	-
Super Vector [27]	-	-	**64.0**	-
Shape alone	15.3	69.0	59.9	21.7
ColorSIFT	20.4	77.6	57.4	-
This paper (universal)	26.3	79.4	61.7	26.5
This paper (specific)	**26.7**	81.3	62.0	**28.1**

Table 2. Comparison of state-of-the-art results with our approach. Note that our approach provides best results on two datasets. The results in the upper part of the table are obtained from the corresponding papers, the results in the bottom part of the table are obtained based on the same detected features.

Finally, we have included the challenging Stanford Dogs 120 dataset. This data set is interesting because dog furs only exist in a reduced set of colors (mainly browns, black and white). Here our approach provides a classification score of 28.1 compared to 21.1 using shape alone. On this dataset, we use the shape features kindly provided by the authors. The universal color descriptor (learned from PASCAL, Birds and Flowers dataset) results in a drop in performance to 26.5. From which we can see that for particular (in a color sense) data sets computing a specific color representation can still yield a large performance gain. To the best of our knowledge the final score of 28.1 obtained in this paper is the best performance achieved on this dataset in literature [4, 3, 17].

Acknowledgments: This work has been supported by project TIN2009-14173 of Spanish Ministry of Science, Swedish Foundation for Strategic Research project Collaborative Unmanned Aerial Systems and Explora' Doc Grant of Region Rhône-Alpes, France.

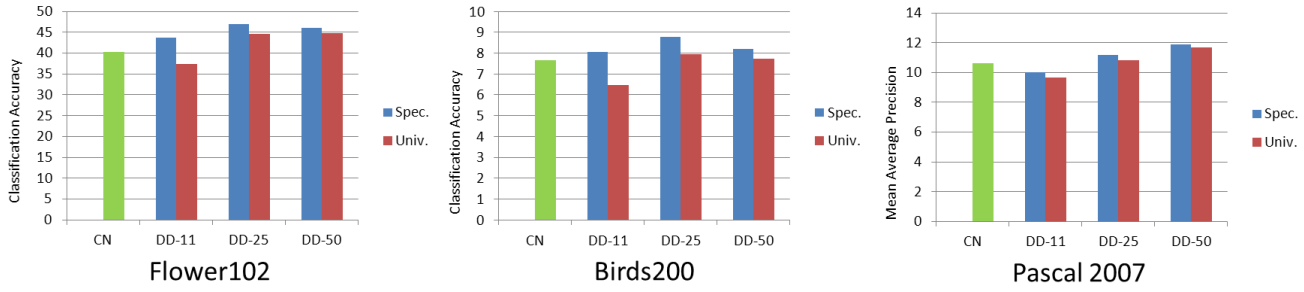


Figure 6. Universality versus Specificity. The green bar (the left bar of each plot) is the state-of-the-art pure color descriptor (Color Names).

References

[1] R. Benavente, M. Vanrell, and R. Baldrich. Parametric fuzzy sets for automatic color naming. *Journal of the Optical Society of America*, 25(10):2582–2593, 2008. 5

[2] B. Berlin and P. Kay. *Basic color terms: their universality and evolution*. Berkeley: University of California, 1969. 5

[3] Y. Chai, V. S. Lempitsky, and A. Zisserman. Bicos: A bi-level co-segmentation method for image classification. In *CVPR*, pages 2579–2586, 2012. 7

[4] Y. Chai, E. Rahtu, V. S. Lempitsky, L. J. V. Gool, and A. Zisserman. Tricos: A tri-level class-discriminative co-segmentation method for image classification. In *ECCV*, 2012. 7

[5] K. Chatfield, V. Lemtexpitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, pages 76.1–76.12, 2011. 7

[6] I. Dhillon, S. Mallela, and R. Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *JMLR*, 3:1265–1287, 2003. 2, 3

[7] N. Elfiky, F. S. Khan, J. van de Weijer, and J. Gonzalez. Discriminative compact pyramids for object and scene recognition. *Pattern Recognition*, 45(4):1627–1636, April 2012. 6

[8] G. Finlayson and S. Hordley. Gamut constrained illumination estimation. *International Journal of Computer Vision*, 67(1):93–109, 2006. 1

[9] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, October 2009. 6

[10] B. Funt and G. Finlayson. Color constant color indexing. *IEEE PAMI*, 17(5):522–529, 1995. 2

[11] J. Geusebroek, R. van den Boomgaard, A. Smeulders, and H. Geerts. Color invariance. *IEEE PAMI*, 23(12):1338–1350, 2001. 2

[12] T. Gevers and A. Smeulders. Color based object recognition. *Pattern Recognition*, 32:453–464, 1999. 1, 2

[13] T. Gevers and H. Stokman. Robust histogram construction from colour invariants for object recognition. *IEEE PAMI*, 26(1):113–118, 2004. 2

[14] F. Khan, R. Anwer, J. van de Weijer, A. Bagdanov, M. Vanrell, and A. Lopez. Color attributes for object detection. In *CVPR*, 2012. 2, 6

[15] F. Khan, J. Van de Weijer, A. Bagdanov, and M. Vanrell. Portmanteau vocabularies for multi-cue image representation. In *Twenty-Fifth Annual Conference on Neural Information Processing Systems (NIPS 2011)*, 2011. 3, 7

[16] F. S. Khan, J. van de Weijer, and M. Vanrell. Modulating shape features by color attention for object recognition. *International Journal of Computer Vision (IJCV)*, 98(1):49–64, 2012. 1, 2, 6, 7

[17] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, CVPR*, Colorado Springs, CO, June 2011. 7

[18] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006. 7

[19] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. 2, 6, 7

[20] S. Shafer. Using color to seperate reflection components. *COLOR research and application*, 10(4):210–218, Winter 1985. 1, 2

[21] A. Torralba and A. Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528. IEEE, 2011. 6

[22] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE PAMI*, 32(9):1582–1596, 2010. 1, 7

[23] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus. Learning color names for real-world applications. *IEEE Transactions on Image Processing*, 18(7):1512–1524, july 2009. 1, 5, 6

[24] E. Vazquez, R. Baldrich, J. van de Weijer, and M. Vanrell. Describing reflectances for color segmentation robust to shadows, highlights, and textures. *IEEE PAMI*, 33(5):917–930, 2011. 3

[25] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 6

[26] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object catergories: A comprehensive study. *IJCV*, 73(2):213–218, 2007. 7

[27] X. Zhou, K. Yu, T. Zhang, and T. Huang. Image classification using super-vector coding of local image descriptors. In *ECCV*, 2010. 7