Color Attributes for Object Detection

Fahad Shahbaz Khan[1,*], Rao Muhammad Anwer[2,*], Joost van de Weijer[2],
Andrew D. Bagdanov[2,3], Maria Vanrell[2], Antonio M. Lopez[2]

[1] Computer Vision Laboratory, Linköping University, Sweden

[2] Computer Vision Center, CS Dept. Universitat Autonoma de Barcelona, Spain

[3] Media Integration and Communication Center, University of Florence, Italy

Abstract

State-of-the-art object detectors typically use shape information as a low level feature representation to capture the local structure of an object. This paper shows that early fusion of shape and color, as is popular in image classification, leads to a significant drop in performance for object detection. Moreover, such approaches also yields suboptimal results for object categories with varying importance of color and shape.

In this paper we propose the use of color attributes as an explicit color representation for object detection. Color attributes are compact, computationally efficient, and when combined with traditional shape features provide state-of-the-art results for object detection. Our method is tested on the PASCAL VOC 2007 and 2009 datasets and results clearly show that our method improves over state-of-the-art techniques despite its simplicity. We also introduce a new dataset consisting of cartoon character images in which color plays a pivotal role. On this dataset, our approach yields a significant gain of 14% *in mean AP over conventional state-of-the-art methods.*

1. Introduction

Object detection is one of the most challenging problems in computer vision. It is difficult due to the significant amount of variation between images belonging to the same object category. Other factors, such as changes in viewpoint and scale, illumination, partial occlusions and multiple instances further complicate the problem of object detection [5, 8, 18, 10, 21, 23]. Most state-of-the-art approaches to object detection rely on intensity-based features that ignore color information in the image [5, 14, 10, 23]. This exclusion of color information is usually due to large variations in color caused by changes in illumination, compression, shadows and highlights, etc. These variations make the task of robust color description especially difficult. On the other hand, and in contrast to object detection, color has been shown to yield excellent results in combination with shape features for image classification [17, 13, 12]. The few approaches which do apply color for object detection focus on a single class such as pedestrians [15, 22, 2]. However, the problem of generic object detection is more challenging and the contribution of color to object detection on standard benchmark datasets such as the PASCAL VOC [8] is yet to be investigated.

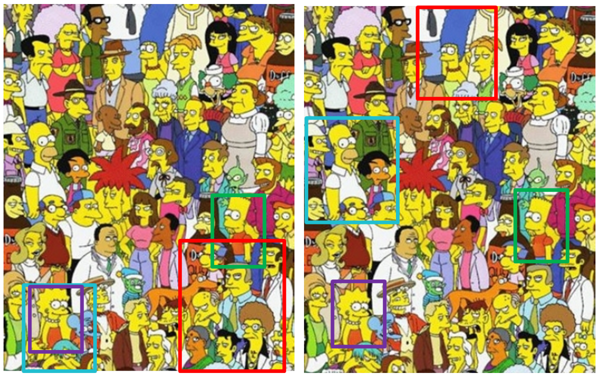

Figure 1. Find the Simpsons. On the left, the conventional part-based approach [10] fails to detect all four members of Simpsons. Only Bart and Lisa are correctly detected, while Homer is falsely detected as Lisa and Marge is not detected at all. On the right, our extension of the part-based detection framework with color attributes can correctly classify all four Simpsons.

In this paper, we investigate extending color information in two existing methods for object detection, specifically the part-based detection framework [10] and the Efficient Subwindow Search approach [14]. The failure of existing approaches motivates us to distinguish three main criteria which should be taken into account when choosing an approach to integrating color into object detection.

Feature Combination: There exist two main approaches to

[*] Both authors contributed equally to this work.

combining shape and color information: early and late fusion [13, 16, 17]. Early fusion combines shape and color at the pixel level, which are then processed together throughout the rest of the learning and classification pipelines [13, 16]. In late fusion, shape and color are described separately from the beginning and the exact binding between the two features is lost. Early fusion, in general, results in more discriminative features than late fusion since it preserves the spatial binding between color and shape features. Due to its high discriminative power, early fusion has traditionally been a very successful tool for image classification [17]. Recent results, however, have shown that when incorporating spatial pyramids, late fusion methods often obtain better results [7]. This is due to the fact that once spatial cells become smaller the uncertainty introduced by describing shape and color separately is reduced. In the limit, where spatial cells represent a single pixel, early and late fusion are equivalent. The importance of smaller cells of spatial pyramids for object detection has been amply demonstrated [11, 5], and therefore our intuition is that late fusion of color and shape will yield better object detection performance than early fusion.

Photometric invariance: One of the main challenges in color representation is the large variation in features caused by scene-accidental effects such as illumination changes and varying shadows. Photometric invariance theory provides guidelines on how to ensure invariance with respect to such events [19, 17], however photometric invariance comes at the cost of discriminative power. The choice of the color descriptor used should take into consideration both its photometric invariance as well as its discriminative power.

Compactness: Existing luminance-based object detection methods use complex representations. For example the part-based method of Felzenswalb [10] models an object as a collection of parts, where each part is represented by a number of histograms of gradient orientations over a number of cells. Each cell is represented by a 31-dimensional vector. Training such a complex model, for just a single class, can require over 3GB of memory and take over 15 hours on a modern, multi-core computer. When extending these cells with color information it is therefore imperative to use a color descriptor as compact as possible both because of memory usage and because of total learning time.

This paper investigates the incorporation of color for object detection based on the above mentioned criteria. We demonstrate the advantages of combining color with shape on the two most popularly used detection frameworks, namely part-based detection with deformable part models [10] and Efficient Subwindow Search (ESS) for object localization [14]. In contrast to conventional fusion approaches that compute shape features on the color channels independently, we propose the use of color attributes as an explicit color representation. The resulting image representations are compact and computationally efficient while providing excellent detection performance on challenging datasets. Figure 1 provides some examples of how our extension correctly detects challenging object classes where state-of-the-art techniques using shape information alone fail.

2. Related work

Most successful approaches to object detection are based on the learning-from-examples paradigm and rely on shape or texture information for image representation [23, 5, 10]. Conventionally, a sliding window approach is used which exhaustively scans an image at multiple locations and scales. An SVM is then trained using positive and negative examples from each object category. Given a test image, a classifier then selects the candidate windows most likely to contain an object instance. Among various features, histograms of oriented gradients (HOG) proposed by Dalal and Triggs [5] are the most commonly used features for object detection.

Recently, discriminative, part-based approaches [23, 10] have been shown to provide excellent performance on the PASCAL VOC datasets [9]. Felzenszwalb et al. [10] propose a star-structured part-based detector where HOGs are used for image representation and latent support vector machines for classification. A boosted HOG-LBP detector is proposed by [23], where LBP descriptors are combined with HOGs to incorporate texture information. A boosting technique is employed for feature selection and their approach yields improved performance for objects. In this paper, we incorporate color information within the part-based framework of Felzenszwalb et al. [10]. Contrary to the approach presented by [23], our approach requires no feature selection to identify relevant features for part representation.

In contrast to part-based detection methods, the bag-of-words model has also been used for object detection [21, 11, 18, 14]. These methods are based on the bag-of-words framework where features are quantized into a visual vocabulary. Vedaldi et al. [21] use a multiple kernel learning framework with powerful visual features for object detection. Harzallah et al. [11] use a two-stage cascade classifier for efficient detection. Their approach also combines object localization and image classification scores. The sliding window approach together with the bag-of-words framework is computationally expensive. Alexe et al. [1] propose an objectness measure to select a few candidate regions likely to contain an object instance in an image. Van de Sande et al. [18] propose the use of hierarchical segmentation as a selective search strategy for object detection. Alternatively, Lampert et el. [14] propose an Efficient Subwindow Search strategy (ESS) to counter the problem of exhaustively scanning sliding windows. In this paper, we also investigate the contribution of color when used in combination with shape features in the ESS framework.

3. Color attributes for object detection

In this section we describe the color descriptors we will use to augment the shape-based feature descriptors used for object detection. Based on the analysis in the introduction section, our approach will apply a late fusion of shape and color.

3.1. Color descriptors

A consequence of our choice of late fusion is that we require a pure color descriptor. In addition, we would like this color descriptor to be discriminative, to possess photometric invariance to some degree, and to be compact. Several color descriptors have been proposed in literature. We consider three of them here.

Robust hue descriptor (HUE) [19]: image patches are represented by a histogram over hue computed from the corresponding RGB values of each pixel according to:

$$hue = \arctan\left(\frac{\sqrt{3}\,(R-G)}{R+G-2B}\right). \tag{1}$$

To counter instabilities in hue, its impact in the histogram is weighted by the saturation of the corresponding pixel. The hue descriptor is invariant with respect to lighting geometry and specularities when assuming white illumination.

Opponent derivative descriptor (OPP) [19]: image patches are represented by a histogram over the opponent angle:

$$ang_{\mathbf{x}}^{O} = \arctan\left(\frac{O1_{\mathbf{x}}}{O2_{\mathbf{x}}}\right) \tag{2}$$

where $O1_{\mathbf{x}}$ and $O2_{\mathbf{x}}$ are the spatial derivatives in the chromatic opponent channels. The opponent angle is weighted by the chromatic derivative strength $\sqrt{O1_{\mathbf{x}}^2 + O2_{\mathbf{x}}^2}$. The opponent angle is invariant with respect to specularities and diffuse lighting.

Color names (CN) [20]: color names, or color attributes, are linguistic color labels which humans assign to colors in the world. Based on several criteria with respect to usage and uniqueness, Berlin and Kay [4] in a linguistic study concluded that the English language contains eleven basic color terms: black, blue, brown, grey, green, orange, pink, purple, red, white and yellow. In computer vision, color attributes involve the assignment of linguistic color labels to pixels in an image. This requires a mapping between RGB values and color attributes [3]. In this paper, we use the mapping learned from Google images in [20] as a color descriptor. Color names display a certain amount of photometric invariance because several shades of a color are mapped to the same color name. They also provide an added advantage of allowing the description of achromatic colors such as black, grey and white which are impossible to distinguish from a photometric invariance perspective. Color names have been found to be a successful color feature for image classification [13]. They have the additional advantage of being a very compact representation.

The color name descriptor is defined as a vector containing the probability of a color name given an image region R:

$$CN = \{p\left(cn_1|R\right), p\left(cn_2|R\right), ..., p\left(cn_{11}|R\right)\} \tag{3}$$

with

$$p\left(cn_i|R\right) = \frac{1}{P}\sum_{x\in R} p\left(cn_i|\mathbf{f}\left(x\right)\right), \tag{4}$$

where cn_i is the i-th color name, x are the spatial coordinates of the P pixels in region R, $\mathbf{f} = \{L^*, a^*, b^*\}$, and $p\left(cn_i|\mathbf{f}\right)$ is the probability of a color name given a pixel value. The probabilities $p\left(cn_i|\mathbf{f}\right)$ are computed from a set of images collected from Google. To learn color names, 100 images per color name are used. To counter the problem of noisy retrieved images, the PLSA approach is used [20].

To summarize, color names possess some degree of photometric invariance. However, they also can encode achromatic colors such as black, grey and white, leading to higher discriminative power.

3.2. Color descriptor evaluation

To select one of the color descriptors described above we performed the following experiment. The histograms of a 2 × 2 spatial pyramid for all the bounding boxes of each object category are extracted. To compare discriminative power, for each histogram we compute KL-ratio between the Kullback-Leibler (KL) divergence of each histogram with members of the other classes and the KL-divergence with members of its own class:

$$\text{KL-ratio} = \frac{\sum\limits_{k\in C^m} \min\limits_{j\notin C^m} KL\left(p_j, p_k\right)}{\sum\limits_{k\in C^m} \min\limits_{i\in C^m, i\neq k} KL\left(p_i, p_k\right)}, \tag{5}$$

where

$$KL\left(p_i, p_j\right) = \sum_{x=1}^{N} p_i\left(x\right)\log\frac{p_i\left(x\right)}{p_j\left(x\right)}, \tag{6}$$

and p_i is the histogram of bounding box i over the N visual words x. Indices $i \in C^m$ represent bounding boxes which belong to class m, while $j \notin C^m$ are random samples of the bounding boxes which are not the same class as m. We choose the number of negative samples j to be three times the size of the positive samples $i \in C^m$. A higher KL-ratio reflects a more discriminative descriptor, since the average intra-class KL-divergence is lower than the inter-class KL-divergence.

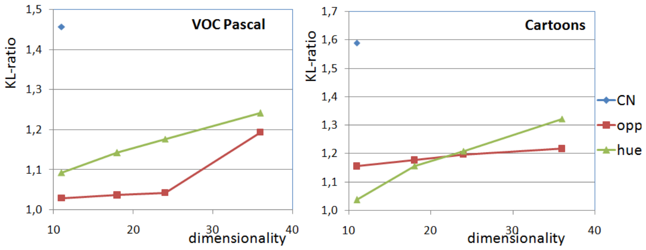

Figure 2. KL-ratio for the PASCAL VOC 2007 and the Cartoon dataset. The graphs clearly show that the color attribute (CN) is superior to the HUE and OPP descriptors in terms of both compactness and discriminative power.

In Figure 2, we report the average KL-ratio over the classes for each color features, HUE, OPP and CN, on the two data sets we use in our experimental evaluation: PASCAL VOC 2007 and a new data set Cartoons. For both HUE and OPP we vary the number of bins in the histogram from 36 as used in [19] to eleven bins which is the size of the CN descriptor. Lowering the dimensionality of the OPP and HUE descriptors leads as expected to lower KL-ratios. As can be seen, the CN descriptor obtains higher KL-ratio even compared to the 36 dimensional representation of the HUE and OPP descriptors.

Based on this experiment we select the CN descriptor as the color feature to use for object detection. It is a pure color feature, therefore allowing us to use it for late fusion with shape features, and based on the KL-ratio it was demonstrated to be more discriminative and compact than the HUE and OPP color descriptors.

4. Coloring object detection

In this section we show how two detection methods can be augmented with color attributes for object detection. We start by coloring a part-based detection framework [10], then we show how color can enhance the performance of ESS-based object localization [14].

4.1. Coloring part-based object detection

In part-based object detection each object is modeled as a deformable collection of parts with a root model at its core [10]. The root filter can be seen as analogous to the HOG-based representation of Dalal and Triggs [5]. Learning in the part-based framework is performed by using a latent SVM formulation. The detection score for a window is obtained by concatenating the root filter, the part filters and the deformation cost of the configuration of all parts. Both the root and the parts are represented by a dense grid of 8x8 non-overlapping cells. A one-dimensional histogram of HOG features is computed over all the pixels in a cell, capturing the local intensity changes.

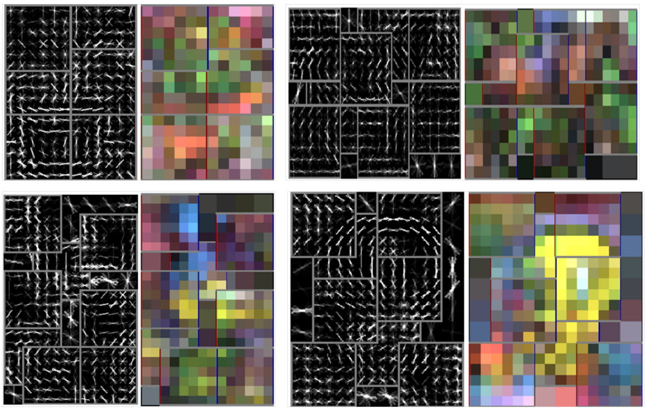

Figure 3. Visualization of learned part-based models with color attributes. Both the HOG and color attribute components of our trained models are shown. Each cell is represented by the color which is obtained by multiplying the SVM weights for the 11 CN bins with a color representative of the color names. Top row: the HOG and color attribute models for pottedplant and horse. Bottom row: Marge and Tweety models. In the case of horse, the brown color of the horse together with a person sitting on top of it is prominent. Similarly, the model is able to capture the blue hair of Marge and orange feet of Tweety.

Conventionally, HOGs are computed densely to represent an image. An image is divided into 8x8 non-overlapping pixel regions known as cells. We follow a similar procedure to compute color attributes for each cell, resulting in a histogram representation. We extend the 31-dimensional HOG vector with the eleven-dimensional color attributes vector. For cell C_i, the representation is obtained by concatenation:

$$C_i = [HOG_i, CN_i]\,, \tag{7}$$

and this concatenated representation thus has dimensionality 42. This is still significantly more compact than an early fusion approach where the HOG would be computed on multiple color channels. Such an approach slows the whole detection pipeline significantly by increasing both time complexity and memory usage. Table 2 shows a comparison of feature dimensions of different extensions of the part-based method.

Throughout the learning of the deformable part-based model both appearance and color are used. Therefore, the introduction of color leads to models which can significantly differ from the models learned on only luminance-based appearance. Examples of four models are provided in Figure 3.

4.2. Coloring ESS object detection

The Efficient Subwindow Search (ESS) object localization framework [14] offers an efficient alternative to the

Feature	HOG	OPPHOG	RGBHOG	C-HOG	LBP-HOG [23]	CN-HOG
Dimension	31	93	93	93	90	42

Table 1. Comparison of feature dimensionality of different approaches. Our proposed CN-HOG feature increases dimensionality to only 42 dimensions. The early fusion extensions of HOG based on computing the HOG on multiple color channels result in dimensionality of 93 (notations are similar to [17]). The LBP-HOG approach combines the LBP and HOG using late fusion and increases overall dimensionality to 90.

computationally expensive bag-of-words approach to sliding window object detection. ESS relies on a branch and bound strategy in order to globally optimize a quality criterion across all sub-windows in an image. ESS is based on a bag-of-words representation of the image. Typically, a number of local features are extracted from each image, and these local features are then quantized into a visual vocabulary from which histograms are generated.

A shape-based visual vocabulary of SIFT features is usually used for detection using the ESS framework [14]. Color can be incorporated using early or late fusion for image representation. Both extensions are straightforward. In early fusion a single combined color-shape vocabulary is created and extracted patches are represented by a color-shape visual word. In late fusion, a separate shape and color vocabulary are learned, and patches are represented by two indexes, one for the shape vocabulary and one for the color vocabulary. Though we use late fusion to incorporate CN features into ESS, we will also compare with ESS results based on early fusion.

5. Cartoon character detection

The PASCAL VOC dataset for object detection is predominantly shape-oriented and color plays a subordinate role [13]. To evaluate the potential contribution of color to object detection, we present a new, publicly available dataset of cartoon character images[1]. The dataset consist of 586 images of 18 popular cartoon characters collected from Google. The 18 cartoon characters in the dataset are: The Simpsons (Bart, Homer, Marge, Lisa), the Flinstones (Fred and Barney), Tom, Jerry, Sylvester, Tweety, Bugs, Daffy, Scooby, Shaggy, Roadrunner, Coyote, Donald Duck and Micky Mouse. The dataset contains a variable number of images for each character, ranging from 28 (Marge) to 85 (Tom). Each class is equally divided into training and testing sets where the number of images per category vary. The dataset is challenging as the images come from sources of different types, such as graphics, wallpapers, sketches, etc. Figure 4 shows some example images from the dataset. Note the variable quality, appearance and scale of the various cartoon characters. To evaluate detection performance,

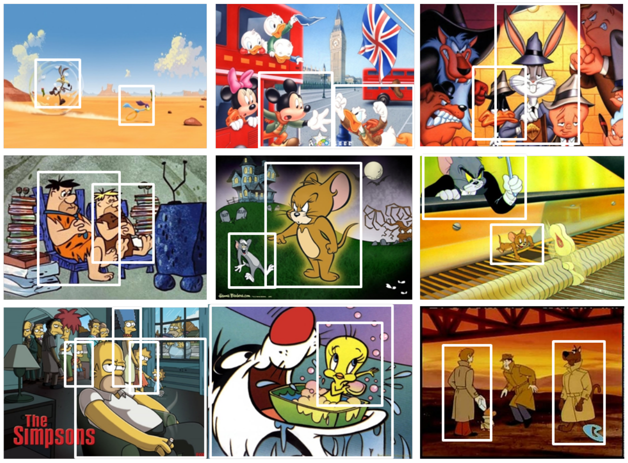

Figure 4. Example images with annotations from the new Cartoon dataset. The dataset consists of images of 18 different cartoon characters.

we follow the PASCAL VOC evaluation criteria [8].

6. Experimental results

Here we first present our results on the PASCAL VOC datasets and in section 6.2 results on the Cartoon dataset.

6.1. Results on the PASCAL VOC datasets

The PASCAL VOC 2007 dataset consists of 9963 images of 20 different object classes with 5011 training images and 4952 test images. The 2009 PASCAL VOC dataset contains 13704 images of 20 different categories. We first provide results of our approach based on the part-based approach [10] discussed in Section 4.1. Afterwards, we show the results obtained using ESS for object detection.

Coloring part-based object detection: The conventional part-based framework [10] is based on HOGs for feature description. We start by comparing our approach with [10] where HOGs with no color information are used as a feature descriptor. We first perform an experiment to compare our proposed approach with existing color descriptors. Recently, a comprehensive evaluation of color descriptors has been presented by Van de Sande et al. [17]. In this evaluation, opponentSIFT and C-SIFT were shown to yield superior performance on image classification.

We also perform experiments using the standard RGB color space. In case of early fusion, HOG features are computed on the color channels and the resulting feature vectors are concatenated in a single representation. To the best of our knowledge, the performance of these color descriptors have not been evaluated before on the task of object detection within a part-based framework. Table 2 shows the results on all 20 categories of the PASCAL VOC 2007 dataset. None of the three color-based methods, namely opponen-

[1] The dataset is available at http://www.cat.uab.cat/Research/object-detection

	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean AP
HOG [10]	28.9	59.5	10.0	15.2	**25.5**	**49.6**	57.9	19.3	**22.4**	**25.2**	23.3	11.1	56.8	48.7	41.9	12.2	17.8	**33.6**	45.1	41.6	32.3
OPPHOG	29.2	54.2	10.7	14.5	17.9	45.8	53.5	21.7	19.3	22.8	21.7	12.3	57.4	46.0	41.2	15.6	19.2	25.0	42.2	41.2	30.6
RGBHOG	33.9	56.5	6.8	13.7	22.9	46.2	56.6	14.9	20.4	22.8	19.3	11.7	57.1	46.7	40.6	13.3	19.2	31.6	47.5	43.4	31.3
C-HOG	29.1	54.7	9.8	14.3	17.9	44.8	55.2	16.0	19.5	25.1	19.6	11.8	58.5	46.6	27.1	15.2	19.0	26.9	44.0	46.6	30.1
CN-HOG (*This paper*)	**34.5**	**61.1**	**11.5**	**19.0**	22.2	46.5	**58.9**	**24.7**	21.7	25.1	**27.1**	**13.0**	**59.7**	**51.6**	**44.0**	**19.2**	**24.4**	33.1	**48.4**	**49.7**	**34.8**

Table 2. Average precision results for the baseline HOG detector [10], color descriptors proposed in the literature [17] and our proposed CN-HOG approach on all 20 classes of the PASCAL VOC 2007 dataset. Note that our approach along among existing fusion methods outperforms shape alone on this dataset. Our approach provides a significant improvement of 2.5% mean AP over the standard HOG-based framework.

tHOG, RGBHOG and C-HOG, improve the performance over the standard HOG features. Our proposed approach, which has the additional advantage of being compact and computationally efficient, results in a significant improvement of 2.5% on the mean average precision over the baseline HOG. Figure 5 shows the precision/recall curves for these color descriptors as well as the baseline HOG on six different object categories from the PASCAL VOC 2007 dataset.

Finally, Table 4 shows results obtained on the PASCAL VOC 2009 dataset. Our proposed approach obtains a gain of 1.4% in mean AP over standard HOG based framework. Our method provides superior results on 15 out of 20 object categories compared to standard HOG based framework. Moreover, independently weighting the contribution of color and shape is expected to improve on categories where adding color provides inferior performance.

Coloring ESS-based object detection: The ESS-based approach has been shown to provide good localization results on the cat and dog categories of the PASCAL VOC 2007 dataset. We only report the results on cat and dog categories since ESS results in similar or better results compared to part-based methods on these two classes. To evaluate the performance of our proposed approach, we construct a 4000 visual word shape vocabulary based on SIFT features. A visual vocabulary of 500 color-words is constructed using the CN descriptor described above.

On the cat category, our proposed approach provides an AP of 22.3% compared to 20.7% obtained using shape alone. Similar results are obtained on the dog category where shape alone and our approach provide score of 13.8 and 15.8 respectively.

Comparison with state-of-the-art results: Table 3 shows a comparison of our approach with the state-of-the-art results reported in literature. Firstly, our proposed color attribute-based approach improves the baseline part-based approach on 15 out of the 20 object categories. The results reported by [21] are obtained by using the bag-of-words framework with multiple features combined using a multiple kernel learning framework. The boosted HOG-LBP approach [23] combines HOG and LBP features while employing boosting as a feature selection mechanism. It is further reported by [23] that without this feature selection strategy, the naive feature combination provides inferior results. In contrast to these approaches, no feature selection strategy is used in our approach, though a selection strategy can be easily incorporated which is expected to further improve results. The approach proposed by [18] provides a mean AP of 33.9% using multiple color spaces, specifically RGB, opponent, normalized rgb and hue for segmentation. Moreover, a dense representation based on SIFT, opponentSIFT and RGBSIFT is used within the bag-of-words framework. Our approach provides the best mean AP reported on this dataset in the detection literature[2] [10, 6, 24, 9, 23, 18]. Finally, on the PASCAL VOC 2009 dataset our approach provides best results on 7 object categories.

6.2. Results on the Cartoon dataset

Here we report the results obtained on our new Cartoon dataset in which color plays an important role. We first show results using the part-based framework and then follow with a comparison of several approaches using the ESS-based object detection framework.

Coloring part-based object detection: Table 5 shows the results obtained using the part-based framework. The conventional part-based approach using HOG features provides a mean AP of 27.6%. The early fusion based approaches yield similar results[3]. Our approach, however, results in a significant gain of 14% in mean AP compared to standard HOG. Moreover, our approach gives the best performance on 9 out of the 18 cartoon categories compared to HOG, opponentHOG, C-HOG and RGBHOG. On categories such as Daffy and Tom, our approach results in a gain of 25.9% and 9.2%, respectively, compared to the second-best approach. This significant gain can credited to the fact that color names have the additional capability of encoding achromatic colors such as black, grey and white.

Coloring ESS-based object detection: Here we compare our approach with shape alone and the two best color descriptors reported in the literature, namely opponentSIFT and C-SIFT [17]. We perform an experiment to compare the performance of different fusion approaches. We use vi-

[2] We do not compare our results with methods combining image classification and detection. Such approaches can be seen as complementary to our approach.

[3] We also performed an experiment with a 36-dimensional hue-saturation color descriptor concatenated with a HOG. This yielded a MAP of 34.2%, significantly lower than the 41.7% of our compact representation.

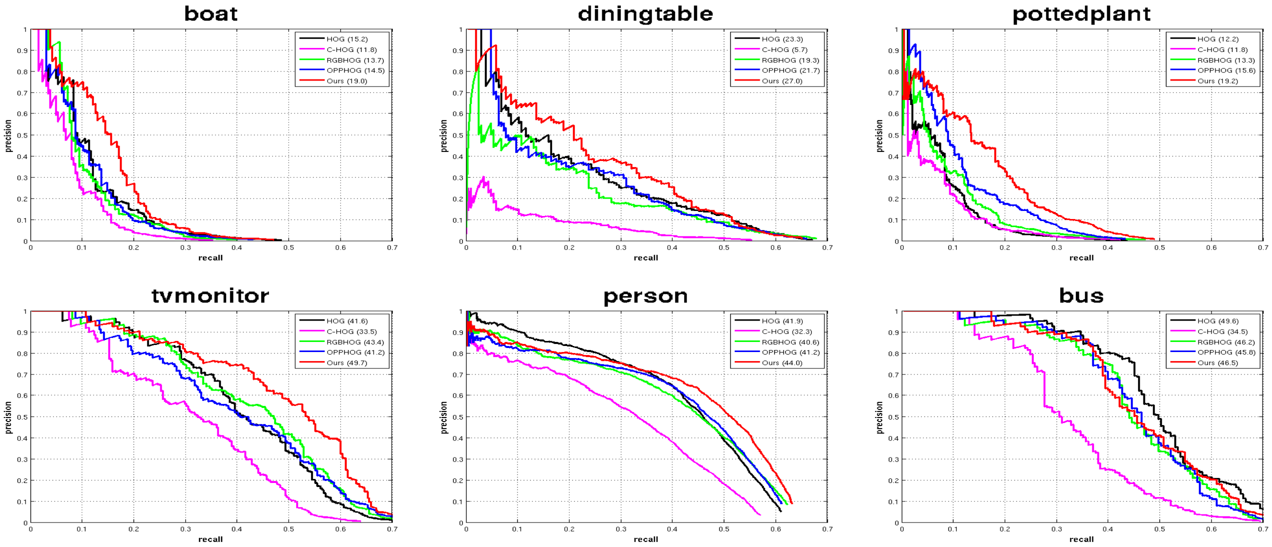

Figure 5. Precision/recall curves of the various approaches on six different categories from the PASCAL VOC 2007 dataset. Other than the bus category, our approach provides significantly improved performance compared to others.

	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean AP
HOG [10]	28.9	59.5	10.0	15.2	25.5	49.6	57.9	19.3	22.4	25.2	23.3	11.1	56.8	48.7	41.9	12.2	17.8	33.6	45.1	41.6	32.3
Best 2007 [9]	26.2	40.9	9.8	9.4	21.4	39.3	43.2	24.0	12.8	14.0	9.8	16.2	33.5	37.5	22.1	12.0	17.5	14.7	33.4	28.9	23.3
UCI [6]	28.8	56.2	3.2	14.2	**29.4**	38.7	48.7	12.4	16.0	17.7	24.0	11.7	45.0	39.4	35.5	15.2	16.1	20.1	34.2	35.4	27.1
LEO [24]	29.4	55.8	9.4	14.3	28.6	44.0	51.3	21.3	20.0	19.3	25.2	12.5	50.4	38.4	36.6	15.1	19.7	25.1	36.8	39.3	29.6
Oxford-MKL [21]	**37.6**	47.8	**15.3**	15.3	21.9	**50.7**	50.6	**30.0**	17.3	**33.0**	22.5	**21.5**	51.2	45.5	23.3	12.4	23.9	28.5	45.3	48.5	32.1
LBP-HOG [23]	36.7	59.8	11.8	17.5	26.3	49.8	58.2	24.0	**22.9**	27.0	24.3	15.2	58.2	49.2	**44.6**	13.5	21.4	**34.9**	47.5	42.3	34.3
CN-HOG (*This paper*)	34.5	**61.1**	11.5	**19.0**	22.2	46.5	**58.9**	24.7	21.7	25.1	**27.1**	13.0	**59.7**	**51.6**	44.0	**19.2**	**24.4**	33.1	**48.4**	**49.7**	**34.8**

Table 3. Comparison with state-of-the-art results on the PASCAL VOC 2007 dataset. Note that the approach of boosted LBP-HOG [23] combines HOG and LBP together using a boosting strategy for feature selection. However, our proposed combination of color names and HOGs (CN-HOG) is compact, computationally inexpensive and uses no feature selection.

sual vocabularies of 500 shape words and 100 color names. For opponentSIFT and C-SIFT, a visual vocabulary of 500 words is constructed. A comparison of different approaches using the ESS framework is shown in table 6. On this dataset approaches based on color-shape fusion improve performance over standard ESS using SIFT alone.

Table 6 shows the comparative performance of our approach. The results obtained using the ESS-based framework is inferior to that obtained using the part-based method. Both opponentSIFT and C-SIFT yield improved results compared to shape alone. Our approach using the ESS framework gives the best performance on the Sylvester category. Interestingly, on the Cartoon dataset ESS again achieves the best results on cats and dogs.

7. Conclusions

We investigate the problem of incorporating color for object detection. Most state-of-the-art object detectors rely on shape while ignoring color. Recent approaches to augmenting intensity-based detectors with color often provide inferior results for object categories with varying importance of color and shape. We propose the use of color attributes as an explicit color representation for object detection. Color attributes are compact, computationally efficient, and possess some degree of photometric invariance while maintaining discriminative power. We show that our approach can significantly improve detection performance on the challenging PASCAL VOC datasets where existing color-based fusion approaches have shown to provide below-expected results. Finally, we introduce a new dataset of cartoon characters where color plays an important role.

Acknowledgments: We acknowledge EU project ERG-TS-VICI-224737; Spanish Research Program Consolider-Ingenio 2010: MIPRCV (CSD200700018); Tuscan Regional project MNEMOSYNE (POR-FSE 2007-2013, A.IV-OB.2); TRA2011-29454-C03-01, Swedish Foundation for Strategic Research project Collaborative Unmanned Aerial Systems and Spanish projects TIN2009-14173, TIN2010-21771-C02-1. Joost van de Weijer acknowledges support of Ramon y Cajal fellowship.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, 2010. 2

[2] R. M. Anwer, D. Vazquez, and A. M. Lopez. Color contribution to part-based person detection in different types of

	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean AP
HOG [10]	38.0	**47.2**	11.3	12.5	28.1	40.5	38.4	23.4	**17.3**	20.6	15.7	13.8	41.0	43.2	41.7	12.8	24.4	16.6	42.4	31.9	28.0
Oxford-MKL [21]	**47.8**	39.8	**17.4**	**15.8**	21.9	42.9	27.7	**30.5**	14.6	20.6	**22.3**	**17.0**	34.6	43.7	21.6	10.2	**25.1**	16.6	**46.3**	**37.6**	27.7
UOCTTI	39.5	46.8	13.5	15.0	**28.5**	**43.8**	37.2	20.7	14.9	22.8	8.7	14.4	38.0	42.0	41.5	12.6	24.2	15.8	43.9	33.5	27.9
HOG-LBP	11.4	27.5	6.0	11.1	27.0	38.8	33.7	25.2	15.0	14.4	16.9	15.1	36.3	40.9	37.0	13.2	22.8	9.6	3.5	32.1	21.9
CN-HOG (*This paper*)	36.3	46.6	12.5	15.2	27.8	43.5	**39.0**	26.3	16.8	**23.2**	18.8	15.0	**41.4**	**46.7**	**43.3**	**14.7**	23.0	18.3	43.6	35.5	**29.4**

Table 4. Average precision results for the baseline HOG detector [10], our proposed CN-HOG approach and state-of-the-art results on all 20 classes of the PASCAL VOC 2009 dataset. Note that our approach provides an improvement of 1.4 mean AP over the standard HOG-based framework.

	bart	homer	marge	lisa	fred	barney	tom	jerry	sylvester	tweety	buggs	daffy	scooby	shaggy	roadrunner	coyote	donaldduck	mickymouse	mean AP
HOG	72.5	47.7	22.7	82.3	54.5	48.8	6.0	20.5	28.3	18.1	**25.2**	5.5	14.6	12.0	3.5	2.4	11.1	20.7	27.6
OPPHOG	**74.2**	37.6	43.2	86.2	61.6	51.2	13.4	30.7	15.8	**54.4**	25.0	8.8	16.2	21.4	**32.5**	10.9	17.1	16.0	34.2
RGBHOG	71.1	**60.6**	33.0	84.7	62.5	47.6	13.0	34.3	**36.1**	41.9	21.7	6.50	11.5	16.6	29.2	**11.1**	13.1	40.2	35.3
C-HOG	65.5	46.7	26.4	84.1	62.4	**60.4**	23.6	38.7	32.8	33.2	20.6	9.7	**23.4**	25.1	29.3	4.6	21.3	25.9	35.2
CN-HOG (*This paper*)	72.3	40.4	**43.4**	**89.8**	**72.8**	55.1	**32.8**	**52.3**	32.9	51.4	22.2	**35.6**	19.8	**25.2**	21.9	10.0	**27.9**	**45.3**	**41.7**

Table 5. Comparison of different fusion approaches using the part-based approach on the Cartoon dataset. Our CN-HOG approach yields a significant gain of 14% in mean AP over the standard HOG-based approach. Compared to early fusion approaches, our approach results in a gain of 6.4%.

	bart	homer	marge	lisa	fred	barney	tom	jerry	sylvester	tweety	buggs	daffy	scooby	shaggy	roadrunner	coyote	donaldduck	mickymouse	mean AP
Shape	19.0	7.9	8.8	23.6	4.2	1.6	17.5	0.01	24.4	0.2	10	9.8	5.0	2.1	5.8	**7.4**	5.4	5.5	8.8
CSIFT	16.3	4.4	**17.2**	17.7	5.7	3.8	20.0	0.8	24.9	1.9	**16.2**	10.8	**7.4**	8.1	**15.3**	5.6	1.9	7.0	10.3
OPPSIFT	12.8	5.7	14.8	**28.6**	6.3	0.8	24.4	0.6	18.5	**5.1**	6.5	3.7	3.6	**12.1**	1.2	4.3	5.2	**12.5**	9.3
CN-SIFT (*This paper*)	**28.7**	**13.6**	9.9	19.2	**18.2**	**5.1**	**24.5**	**1.9**	**37.4**	0.1	9.9	**13.3**	7.1	11.7	14.0	3.5	**9.8**	4.0	**12.9**

Table 6. Comparison of different approaches within the ESS detection framework. Similar to our results using the part-based method, combining color attributes improves the overall performance by 4%. Our CN-SIFT method also yields superior performance compared to the well known color descriptors OpponentSIFT and C-SIFT.

scenarios. In *CAIP*, 2011. 1

[3] R. Benavente, M. Vanrell, and R. Baldrich. Parametric fuzzy sets for automatic color naming. *JOSA*, 25(10):2582–2593, 2008. 3

[4] B. Berlin and P. Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley, CA, 1969. 3

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1, 2, 4

[6] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009. 6, 7

[7] N. Elfiky, F. S. Khan, J. van de Weijer, and J. Gonzalez. Discriminative compact pyramids for object and scene recognition. *Pattern Recognition Journal*, 2011. 2

[8] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 1, 5

[9] M. Everingham, L. V. Gool, C. K. I.Williams, J.Winn, and A. Zisserman. The PASCAL Visual Object Classes challenge 2007 results. 2, 6, 7

[10] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010. 1, 2, 4, 5, 6, 7, 8

[11] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *ICCV*, 2009. 2

[12] F. S. Khan, J. van de Weijer, A. D. Bagdanov, and M. Vanrell. Portmanteau vocabularies for multi-cue image representations. In *NIPS*, 2011. 1

[13] F. S. Khan, J. van de Weijer, and M. Vanrell. Modulating shape features by color attention for object recognition. *IJCV*, 2011. 1, 2, 3, 5

[14] C. Lampert, M. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*, 2008. 1, 2, 4, 5

[15] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis. Human detection using partial least squares analysis. In *ICCV*, 2009. 1

[16] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *ACM MM*, 2005. 2

[17] K. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 32(9):1582–1596, 2010. 1, 2, 5, 6

[18] K. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011. 1, 2, 6

[19] J. van de Weijer and C. Schmid. Coloring local feature extraction. In *ECCV*, 2006. 2, 3, 4

[20] J. van de Weijer, C. Schmid, J. J. Verbeek, and D. Larlus. Learning color names for real-world applications. *TIP*, 18(7):1512–1524, 2009. 3

[21] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *ICCV*, 2009. 1, 2, 6, 7, 8

[22] S. Walk, N. Majer, K. Schindler, and B. Schiele. New features and insights for pedestrian detection. In *CVPR*, 2010. 1

[23] J. Zhang, K. Huang, Y. Yu, and T. Tan. Boosted local structured hog-lbp for object localization. In *CVPR*, 2010. 1, 2, 5, 6, 7

[24] L. Zhu, Y. Chen, A. L. Yuille, and W. T. Freeman. Latent hierarchical structural learning for object detection. In *CVPR*, 2010. 6, 7