Top-Down Color Attention for Object Recognition

Fahad Shahbaz Khan, Joost van de Weijer, Maria Vanrell
Centre de Visio per Computador, Computer Science Department
Universitat Autonoma de Barcelona, Edifci O, Campus UAB (Bellaterra), C.P.08193, Barcelona,Spain
{ fahad, joost, maria.vanrell } @cvc.uab.cat

Abstract

Generally the bag-of-words based image representation follows a bottom-up paradigm. The subsequent stages of the process: feature detection, feature description, vocabulary construction and image representation are performed independent of the intentioned object classes to be detected. In such a framework, combining multiple cues such as shape and color often provides below-expected results.

This paper presents a novel method for recognizing object categories when using multiple cues by separating the shape and color cue. Color is used to guide attention by means of a top-down category-specific attention map. The color attention map is then further deployed to modulate the shape features by taking more features from regions within an image that are likely to contain an object instance. This procedure leads to a category-specific image histogram representation for each category. Furthermore, we argue that the method combines the advantages of both early and late fusion.

We compare our approach with existing methods that combine color and shape cues on three data sets containing varied importance of both cues, namely, Soccer (color predominance), Flower (color and shape parity), and PASCAL VOC Challenge 2007 (shape predominance). The experiments clearly demonstrate that in all three data sets our proposed framework significantly outperforms the state-of-the-art methods for combining color and shape information.

1. Introduction

Images play a crucial role in our daily communication and the huge amount of pictures digitally available on the internet are not manageable by humans anymore. However, automatic image concept classification is a difficult task, due to large variations between images belonging to the same class. Several other factors such as significant variations in viewpoint and scale, illumination, partial occlusions, multiple instances, also have a significant influence on the final results and thus make the problem of description of images even more complicated. The bag-of-visual words framework, where images are represented by a histogram over visual words, is currently one of the most successful approaches to object recognition. Many features such as color, texture, shape, and motion have been used to describe visual information for object recognition. In this paper, we analyze the problem of object recognition within the bag-of-words framework using multiple cues, particularly, combining shape and color information.

In order to combine multiple cues within the bag-of-words framework, we consider two properties that are especially desirable for the final image representation: *feature binding* and *vocabulary compactness*. Feature binding involves combining information from different features at the local level and *not* at the image level. This allows to differ images with red circles and green squares, from images with green circles and red square. Vocabulary compactness denotes to the property of having a separate visual vocabulary for each of the different cues. This prevents the different cues from getting diluted, which happens in case of a combined shape-color vocabulary. For example, when learning the concept “square” the color cue is irrelevant. In this case, a combined color-shape vocabulary only complicates the learning of “squares” since they are spread across multiple visual words (i.e. blue square, green square, red square etc.). Existing approaches [3, 28, 27, 23] to combine color and shape information into a single framework have not succeeded so far in combining both these properties for image representation.

Conventionally, the bag-of-words based image representation follows a bottom-up paradigm. The subsequent stages of the process: feature detection, feature description, vocabulary construction and image representation are performed independent of the intentioned object categories to be detected. To obtain our goal of combining the two properties discussed above for image representation, we propose to introduce top-down information at an early stage (see Fig. 1). We separate the two cues into a bottom-up *descriptor cue*, in our case shape, and a top-down *attention cue*, color. To this end, we shall use learned class-specific

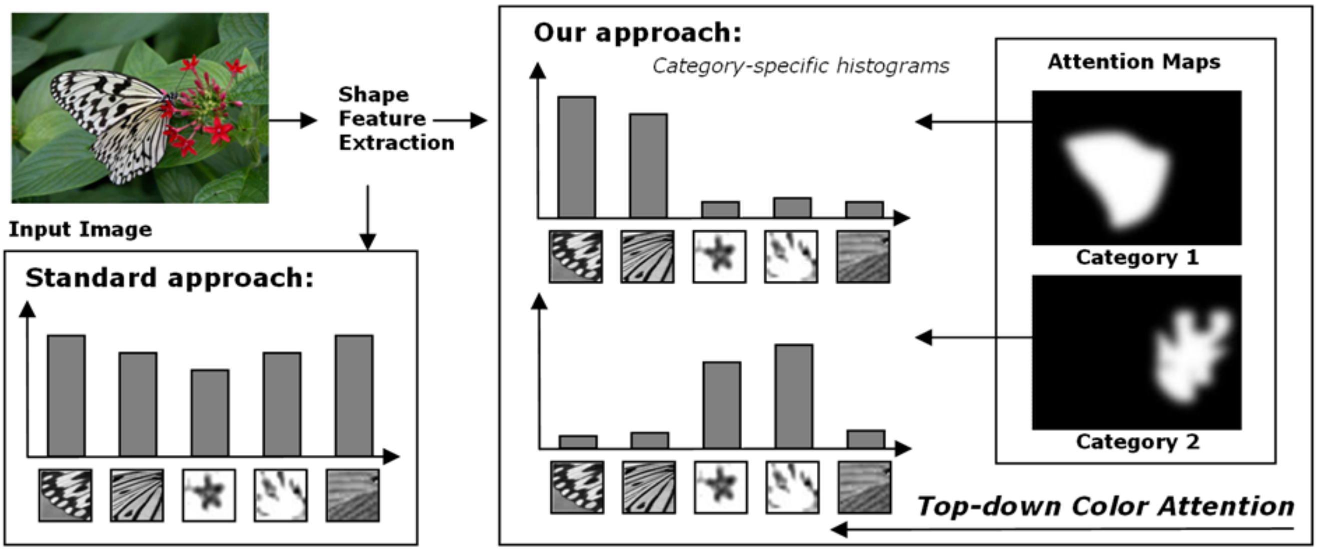

Figure 1. A brief, concise overview of our method. In the standard bag-of-words image representation, the histogram over visual shape words, is constructed in a bottom-up fashion. In our approach we use top-down category-specific color attention to modulate the impact of the shape-words in the image during the histogram construction. Consequently, a separate histogram is constructed for the all categories, where the visual words relevant to each category (in this case flowers and butterflies) are accentuated.

color information to construct a visual attention map of the categories. Subsequently, this color attention map is used to modulate the sampling of the shape features. In regions with higher attention more shape feature are sampled than in regions with low attention. As a result a class-specific image histogram is constructed for each category. We shall show that this image representation combines both properties, feature binding and vocabulary compactness, for image representation. We therefore expect to obtain a superior performance gain from the combination of shape and color, than methods which only possess one of these properties.

The paper is organized as follows. In Section 2 we discuss related work. In Section 3 our approach is outlined based on an analysis of the relative merits of early and late fusion techniques for combining color and shape. In Section 4 the experimental setup is explained. In Section 5, experimental results are given, and Section 6 finishes with concluding remarks.

2. Related Work

There has been a large amount of success in using bag-of-visual-words models for object and scene classification [3, 5, 7, 12, 16, 22, 28] due to its simplicity and very good performance. The first stage in the method involves selecting keypoints or regions followed by representation of these keypoints using local descriptors. The descriptors are then vector quantized into a fixed-size codebook. Finally, the image is represented by a histogram over the visual code-book. A classifier is then trained to recognize the categories based on these histogram representations of the images. Within the bag-of-words framework the optimal fusion of different cues, such as shape, texture and color, still remains open to debate.

Initially, many methods only used the shape features, predominantly represented by SIFT [13] to represent an image [5, 7, 12]. However, more recently the possibility of adding color information has been investigated. Bosch et al. [3] propose to compute the SIFT descriptor in the HSV color space and concatenate the results into one combined color-shape descriptor. Van de Weijer and Schmid [28] compare photometrically invariant representations in combination with SIFT for image classification. Van de Sande et al. [27] performed a study into the photometric properties of many color descriptors, and did an extensive performance evaluation.

There exist two main approaches to fuse color and shape into the bag-of-words representation. The first approach, called *early fusion* involves fusing local descriptors together and creating one joint shape-color vocabulary. The second approach, called *late fusion* concatenates histogram representation of both color and shape, obtained independently. Most of the existing methods use early fusion [3, 27, 28]. Previous work which compares both early and late fusion schemes for image classification have been done by [23, 10] where both early fusion and late fusion are compared. The comparison performed in both studies suggests that combining multiple cues usually improves final classification results but the optimal fusion scheme is still uncertain.

Introducing top-down information into earlier stages of the bag-of-words approach has been pursued in various previous works as well, especially in the vocabulary construction phase. Lazebnik and Raginsky [11] propose to learn discriminative visual vocabularies. The vocabulary construction is optimized to separate the class labels. Perronnin [21] proposes to learn class-specific vocabularies. The image is represented by one universal vocabulary and one adaptation of the universal vocabulary for each of the classes. Both methods showed to improve bag-of-words representations, but they do not handle the issue of multiple cues, and for this reason could be used in complement with the approach presented here. Vogel and Schiele [31] semantically label local features into a number of semantic concepts for the task of scene classification.

The human visual system performs an effective attention mechanism, employed to reduce the computational cost of a data-driven visual search [26, 4]. The higher level vision tasks, such as object recognition, can then focus on the interesting parts in an image to robustly recognize different object categories. Studies of inattentional blindness [1, 24] have revealed that attention itself and its attributes remain unnoticed unless it is directed towards interesting locations in a scene. Thus the visual system selects only a subset of available information and demotes the rest of the information to only a limited analysis. Most natural searches involve targets that are defined by basic feature information. These visual features are loosely bundled into objects before the arrival of attention. In order to bind these features into a recognizable object, attention is required [34].

The two ways by which information can be used to direct attention are, bottom-up (memory-free), where the attention is directed to the *salient regions* and, top-down (memory-dependent), which enables *goal directed* visual search [33]. In computer vision, several works focus on computational visual attention, most of them are based on building saliency maps for simple visual tasks as keypoint detection [8]. However, some attempts has been done towards increasing the feedback of top-down processes into the feature vectors [2]. The work presented in our paper utilizes the top-down visual attention mechanism where the goal is to recognize a specific object category.

Among several properties of visual stimuli, only few are used to control the deployment of visual attention [35]. Color is one such attribute which is undoubtedly used to guide visual attention [35]. Jost et al. [9] measures the contribution of chromatic cue in the model of visual attention. Several other studies performed recently also reveal the importance of color in visual memory [25, 32]. Similarly, in our work color plays a twofold role, firstly, it contains some additional information which is category-specific, and secondly, it modulates the shape words which are computed using a standard bag-of-words approach.

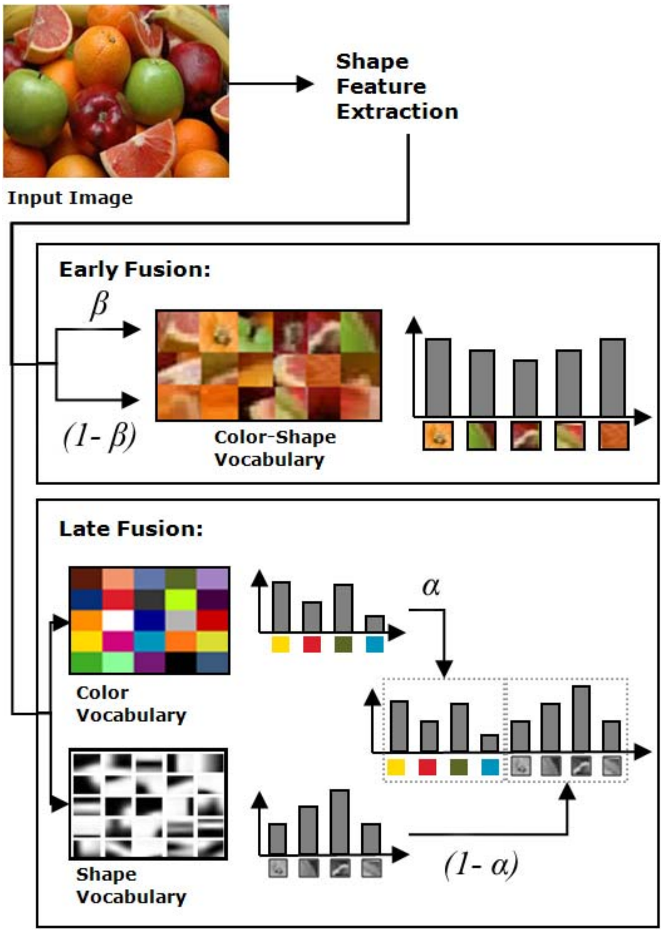

Figure 2. A Graphical explanation of early and late fusion schemes to combine color and shape information. The α and β parameters determine the relative weight of the two cues.

3. Top-Down Color Attention for Object Recognition

In this section, we analyze the two well-known approaches to incorporate multiple cues within the bag-of-words framework, namely early and late fusion. On the basis of this analysis we propose a new approach for the combination of shape and color information for image representation.

3.1. Early and Late Feature Fusion

Before discussing early and late fusion in more detail, we introduce some mathematical notations. In the bag-of-words framework a number of local features f_{ij}, j=1...M^i

are detected in each image I_i, i=1,2,...,N . Generally, the local features are represented in visual vocabularies which describe various image cues such as shape, texture, and color. We shall focus here on shape and color but the theory can easily be extended to include other cues. We assume that visual vocabularies for the cues are available, with the visual words w_i^k, i=1,2,...,V^k and $k \in \{s, c, sc\}$ for the two cues shape and color and for the combined visual vocabulary of color and shape. The local features are represented differently for the two approaches: by a pair of visual words $f_{ij} = \{w_{ij}^s, w_{ij}^c\}$ for late fusion and by single shape-color word $f_{ij} = \{w_{ij}^{sc}\}$ in the case of early fusion.

For a standard single-cue bag-of-words, images are represented by a frequency distribution over the visual words:

$$n\left(\mathrm{w}^{\mathrm{k}}|I^i\right) = \sum_{j=1}^{M^i} \delta\left(w_{ij}^k, \mathrm{w}^{\mathrm{k}}\right) \qquad (1)$$

with

$$\delta\left(x, y\right) \left\{ \begin{array}{ll} 0 & \text{for } x \neq y \\ 1 & \text{for } x == y \end{array} \right. \qquad (2)$$

For early fusion we compute $n\left(w^{\mathrm{sc}}|I^i\right)$. For late fusion we compute $n\left(w^{\mathrm{s}}|I^i\right)$ and $n\left(w^{\mathrm{c}}|I^i\right)$ separately and concatenate the results. A graphical representation of the construction of early and late fusion representations is provided in Fig. 2. The parameters α and β in Fig. 2 indicate the relative weight of color and shape and are learned by cross-validation.

The two approaches early and late fusion lead to different properties for the image representation. We shall discuss the two properties which we believe to be especially important and which form the motivation to our approach. The first property is *vocabulary compactness*. Having this property implies constructing a separate visual vocabulary for both color and shape. This is especially important for classes which only exhibit constancy over one of the cues. For example, many man-made objects are only constant over shape, and vary significantly in color. Learning these classes from a combined shape-color vocabulary only complicates the task of the classifier. Late fusion possesses the property of vocabulary compactness, whereas early fusion lacks it. The second property is *feature binding*. This property refers to methods which combine color and shape information at the local feature level. This allows for the description of blue corners, red blobs, etc. Early fusion has this property since it describes the joined shape-color feature for each local feature. Late fusion, which separates the two cues, only to combine them again at an image-wide level, lacks this property.

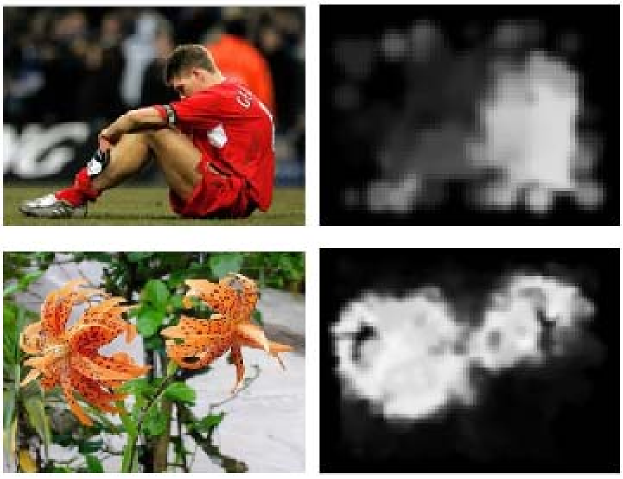

Figure 3. Color attention maps. First row: a liverpool category image from Soccer dataset. Second row: a snowdrop flower species image from Flower dataset.

3.2. Top-Down Color Attention for Image Representation

To obtain an image representation which combines the two desired properties discussed above, we separate the functionality of the two cues. The shape cue will function as a *descriptor cue*, and is used similarly as in the traditional bag-of-words approach. However, the color cue is used as an *attention cue*, and determines the impact of the local features on the image representation. The computation of the image representation is done according to:

$$n\left(\mathrm{w}^{\mathrm{s}}|I^i, class\right) = \sum_{j=1}^{M^i} p\left(class|w_{ij}^c\right) \delta\left(w_{ij}^s, \mathrm{w}^{\mathrm{s}}\right). \qquad (3)$$

The attention cue is used in a top-down manner and describes our prior knowledge about the categories we are looking for. The probabilities $p\left(class|w_{ij}^c\right)$ are computed by using Bayes,

$$p\left(class|\mathrm{w}^{\mathrm{c}}\right) \propto p\left(\mathrm{w}^{\mathrm{c}}|class\right) p\left(class\right) \qquad (4)$$

where $p\left(\mathrm{w}^{\mathrm{c}}|class\right)$ is the empirical distribution,

$$p\left(class|\mathrm{w}^{\mathrm{c}}\right) \propto \sum_{I^{class}} \sum_{j=1}^{M^i} \delta\left(w_{ij}^c, \mathrm{w}^c\right), \qquad (5)$$

obtained by summing over the indexes to the training images of the category I^{class}. The prior over the classes $p\left(class\right)$ is obtained from the training data.

If we compute $p\left(class|w_{ij}^c\right)$ for all local features in an image we can construct a class-specific color attention map. Several examples are given in Fig. 3. The color attention map can be understood to modulate the weighting of the local features. In regions with high attention more shape-features are sampled than in regions with low attentions (note that all histograms are based on the same set of detected features and only the weighting varies). As a consequence a different distribution over the same shape words is obtained for each category, as is indicated in Fig. 1. The final representation of an image is obtained by concatenating the class-specific histograms. Hence, its dimensionality will be equal to the shape vocabulary size V^s times the number of classes.

The image representation proposed in Eq. 3 does not explicitly code the color information. However, indirectly color information is hidden in these representations since the shape-words are weighted by the probability of the category given the corresponding color-word. Some color information is expected to be lost in the process, however the information most relevant to the task of classification is expected to be preserved. Furthermore, this image representation does combine the two properties *feature binding* and *vocabulary compactness*. As can be seen in Eq. 3 the cues are represented in separate vocabularies and are combined at the local feature level. For categories where color is irrelevant, $p\left(class|\mathrm{w}^c\right)$ is uniform and Eq. 3 simplifies to the standard bag-of-words representation of Eq. 1.

The method can simply be extended to include multiple attention cues. For n attention cues we compute

$$n\left(\mathbf{w^s}|I^i, class\right) = \sum_{j=1}^{M^i} p(class|w_{ij}^{c^1}) \times ... \times p(class|w_{ij}^{c^n})\delta\left(w_{ij}^s, \mathrm{w^s}\right) \tag{6}$$

Note that the dimensionality of the image representation is independent of the number of attention cues. Therefore, we also provide results based on multiple color cues. We summerize the procedure of top-down color guided attention image representation in Algorithm 1.

Algorithm 1 Top-Down Color Attention

```
**Require:** Separate visual vocabularies for shape and color with visual words $w_i^k$, i=1,2,...,$V^k$ and $k \in \{s, c\}$ for shape and color.
Construct color histogram $n\left(\mathrm{w^k}|I^i\right)$: images are represented by frequency distribution over color visual words using equation 1.
Compute a class-specific color posterior $p\left(class|w_{ij}^c\right)$: for all the local color features in an image using equation 5.
Construct the class-specific image representation $n\left(\mathrm{w^s}|I^i, class\right)$: compute the weighted class-specific histogram using equation 3.
The final image representation is obtained by concatenating the class posterior $p\left(class|w_{ij}^c\right)$ for all the categories. The dimensionality of the final histogram is $V^s$ times the number of categories.
```

4. Experimental Setup

Details of the proposed procedure are outlined in this section. First, we discuss the implementation details of the descriptors and detectors used for our experiments, followed by a brief description of the data sets used for the evaluation purpose.

4.1. Implementation Details

To test our method, we have used the difference of Gaussian (DoG) detector for the Soccer data set. For Flower data set and PASCAL VOC Challenge 2007 we have used a combination of Harris-Laplace point detector [17] along with DoG and multiscale Grid detector. We normalized all the patches to a standard size and descriptors are computed for all regions in the feature description phase. A visual vocabulary is then computed by clustering the descriptor points using K-means algorithm. In our approach the SIFT descriptor is used to create a shape vocabulary. For color vocabulary we have used two different color descriptors namely, the Color Name (CN) descriptor [29, 30] and HUE descriptor (HUE) [28]. We shall abbreviate our results with the notation convention $CA(descriptor\ cue, attention\ cues)$ where CA stands for Color Attention based bag-of-words. We shall provide results with one attention cue $CA(SIFT,\ HUE)$, $CA(SIFT,\ CN)$, and for the color attention based on two attention cues $CA(SIFT, \{HUE, CN\})$ combined using Eq. 6. Each image is represented by a frequency histogram of visual words. A classifier is then trained based on these histograms. In our experiments we use a standard non-linear SVM with χ^2 kernel for Soccer and Flower data set and intersection kernel for Pascal VOC 2007 data set since it requires significantly less computational time [14], while providing performance similar to χ^2 kernel.

We compare our method with the standard methods to combine color and shape features from literature: early fusion and late fusion. We perform early and late fusion with both CN and HUE and report the best results. Recently, an extensive performance evaluation of color descriptors has been presented by van de Sande et al. [27]. We shall compare our results to the two descriptors reported to be superior. OpponentSIFT uses all the three channels $(O1, O2, O3)$ of the opponent color space. The $O1$ and $O2$ channels describe the color information in an image whereas $O3$ channel contains the intensity information in an image. The WSIFT descriptor is derived from the opponent color space as $\frac{O1}{O3}$ and $\frac{O2}{O3}$, thereby making it invariant with respect to light intensity. Furthermore, it has also been mentioned in [27] that with no prior knowledge about ob-

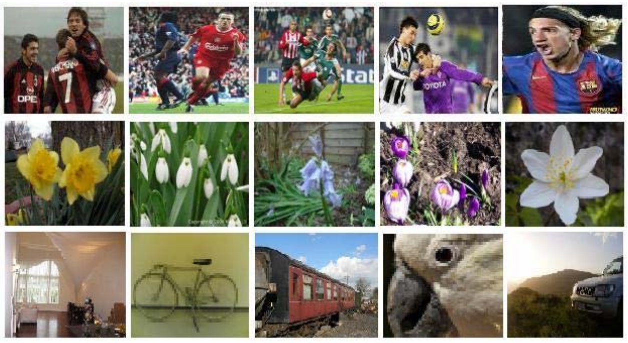

Figure 4. Examples figures of the three data sets. From top to bottom: Soccer, Flower and PASCAL VOC 2007 data set.

ject categories, OpponentSIFT descriptor was found to be the best choice.

4.2. Image Data Sets

We tested our approach on three different and challenging data sets namely Soccer, Flower and PASCAL VOC Challenge 2007. The data sets vary in the relative importance of the two cues shape and color.

The Soccer data set [1] consists of 7 classes of different Soccer teams [28]. Each class contains 40 images divided in 25 train and 15 test images per category. The Flower data set [2] consists of 17 classes of different varieties of flower species and each class has 80 images, divided in 60 train and 20 test images [18]. Finally, we also tested our approach on PASCAL Visual Object Classes Challenge [6]. The PASCAL VOC Challenge 2007 data set [3] consists of 9963 images of 20 different classes with 5011 training images and 4952 test images. Fig. 4 shows some images from the three data sets.

5. Experiments

In this section we present the results of our method on image classification. The data sets have been selected to represent a varied importance of the image cues color and shape. Results are compared to state-of-the-art methods that fuse color and shape cues.

5.1. Image classification: color predominance

Image classification results are computed for the Soccer data set to test color and shape fusion under conditions where color is the predominant cue. In this data set the task is to recognize the Soccer team present in the image. In this case, the color of the player's outfit is the most discriminative feature available.

The results on the Soccer data set are given in Table 1. The importance of color for this data set is demonstrated by the unsatisfactory results of shape alone. The disappointing results for WSIFT might be caused by the importance of the achromatic colors in this data set to recognize the team shirts (for example, Milan outfits are red-black and PSV outfits are red-white). This information might get lost in the photometric invariance of WSIFT. Color Names performed very well here due to their combination of photometric robustness and the ability to describe the achromatic regions. A further performance gain was obtained by combining hue and color name based color attention. Moreover, our approach outperforms the best results reported in literature [29], where a score of 89% is reported, based on a combination of SIFT and CN in an early fusion manner.

Method	Voc Size	Score
$SIFT$	400	50
$EarlyFusion$	1200	90
$LateFusion$	$400 + 300$	90
$WSIFT$	1200	77
$OpponentSIFT$	1200	87
$CA(SIFT, CN)$	$400, 300$	88
$CA(SIFT, HUE)$	$400, 300$	82
$CA(SIFT, \{CN, HUE\})$	$400, \{300, 300\}$	**94**

Table 1. Classification Score (percentage) on Soccer Data set.

5.2. Image Classification: color and shape parity

Image results on the Flower data set show the performance of our approach on a data set for which both shape and color information are vital. The task is to classify the images into 17 categories of flower-species. The use of both color and shape are important as some flowers are clearly distinguished by shape, e.g. daisies and some by color, e.g. fritillaries. The results on Flower data set are given in Table 2.

Method	Voc Size	Score
$SIFT$	1200	68
$EarlyFusion$	2000	85
$LateFusion$	$1200 + 300$	84
$WSIFT$	2000	77
$OpponentSIFT$	2000	83
$CA(SIFT, CN)$	$1200, 300$	87
$CA(SIFT, HUE)$	$1200, 300$	87
$CA(SIFT, \{CN, HUE\})$	$1200, \{300, 300\}$	**89**

Table 2. Classification Score (percentage) on Flower Data set.

Among the existing methods Early Fusion provides the best results. However, the three methods based on color attention obtain significantly better results. Again the combination of CN and HUE was found to give the best results.

[1] The Soccer set at http://lear.inrialpes.fr/data

[2] The Flower set at http://www.robots.ox.ac.uk/vgg/research/flowers/

[3] The PASCAL VOC Challenge 2007 at http://www.pascal-network.org/challenges/VOC/voc2007/

On this data set our method surpassed the best results reported in literature [20]. The best reported result [20] is 88.3% where shape, colour and texture descriptors were computed on the segmentation scheme proposed by [19]. On the other hand neither segmentation nor any bounding box knowledge have been used in our method. A more proximal comparison with our approach is that of [29] where a result of 81 was obtained by combining SIFT and CN in an early fusion manner.

5.3. Image Classification: shape predominance

Finally, we test our approach on a data set where the shape cue is predominant and color plays a subordinate role. The Pascal VOC 2007 challenge data set contains nearly 10,000 images of 20 different object categories. For this data set the average precision is used as a performance metric in order to determine the accuracy of recognition results. The average precision is proportional to the area under a precision-recall curve. The average precisions of the individual classes are used to get a *mean average precision* (MAP) as used by [27]. In the table we have also presented the results in terms of median average precision. Both these statistical metrics are commonly used to evaluate the results on this data set.

Table 3 shows the results. We here only compare against WSIFT which was shown to obtain the best results in [27]. The methods based on color attention again obtain significantly better results for both median and mean AP. For this data set the combination of the two attention cues, HUE and CN, again provides the best results. To obtain state-of-the-art results obtained on PASCAL, the method should be further extended to include spatial information and similarly more complex learning methods should be applied to improve the results further [15].

The results per object category are given in Fig. 5. The 20 categories can be divided into 4 types namely, Animal: bird, cat, cow, dog, horse, sheep, Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train, Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor and Person: person. It is worthy to observe that our approach performs substantially better over all the 20 categories compared to WSIFT and SIFT. Recall that early fusion approaches lack vocabulary compactness and struggle with categories where one cue is constant and the other cue varies a great deal. This behaviour can be observed in vehicle categories such as car, where the color varies significantly over the various instances, something which is known to bother early-fusion methods (i.e. lack of vocabulary compactness). In such classes WSIFT provides below-expected results. Our approach, which combines the advantages of early and late fusion, obtains good results on all types of categories in the data set.

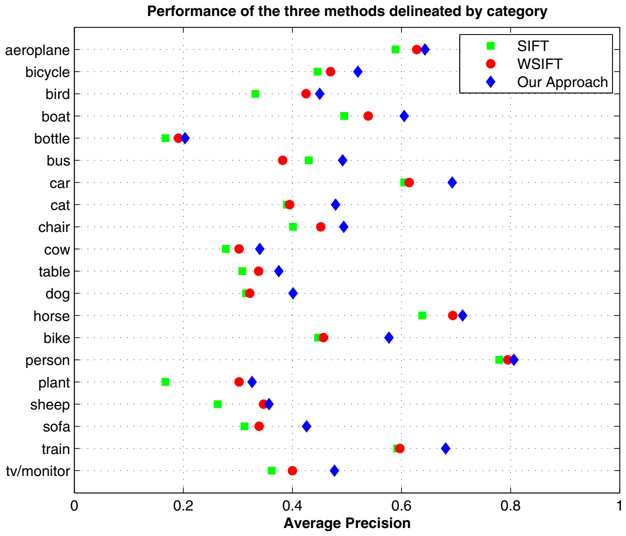

Figure 5. Performance on Pascal VOC Challenge 2007 for SIFT, WSIFT and CA(SIFT,{CN,HUE}). The results are split per object category. Note that we outperform SIFT and WSIFT in all 20 categories.

Method	Voc Size	Median AP	Mean AP
$SIFT$	1000	41.0	43.5
$WSIFT$	6000	41.3	45.0
$CA(SIFT, CN)$	1000, 300	46.4	48.0
$CA(SIFT, HUE)$	1000, 300	48.3	49.5
$CA(SIFT, \{CN, HUE\})$	1000, {300, 300}	**48.6**	**50.2**

Table 3. Median and Mean Average Precision on Pascal VOC Challegenge 2007 data set. Note that our results provide better results both in terms of median and mean AP.

6. Discussion and Conclusions

In this paper we presented a new approach to combine color and shape information within the bag-of-words framework. The methods splits the cues in a bottom-up *descriptor cue* and a top-down *attention cue*. We combine the advantages of early and late fusion, *feature binding* and *vocabulary compactness*, which in a standard bag-of-words approach are mutually exclusive.

The results provided from the three data sets suggest that for most object categories color attention plays a pivotal role in object recognition. Color attention based bag-of-words representations is shown to outperform early and late fusion methods on all three data sets.

It should be noted that our approach is *non-parametric* in that there is no parameter to tune the relative weight of color and shape information (such a parameter is present for both early and late fusion). This could however be easily introduced, for example by

$$n\left(\mathbf{w}^{\mathbf{s}}|I^{i},class\right)=\sum_{j=1}^{M^{i}}p(class|w_{ij}^{c})^{\gamma}\delta\left(w_{ij}^{s},\mathrm{w}^{\mathrm{s}}\right). \quad (7)$$

We expect that learning of such a γ could still further improve the results. A second research direction for future work, will be the inclusion of texture descriptors as attention cues.

Acknowledgements

This work is funded by the Project MEC TIN2007-64577, Ramon y Cajal Program, Consolider-Ingenio 2010 CSD2007-00018, of Spanish Ministry of Science and ERG-TS-VICI-224737 (European Re-integration Grants) funded by Marie-Curie Actions of the EU.

References

[1] A.Mack and I. Rock. *Inattentional blindness*. MIT Press., 1998.

[2] A.Oliva and A.Torralba. Top-down control of visual attention in object detection. In *ICIP*, 2003.

[3] A. Bosch, A. Zisserman, and J.Munoz. Scene classification via plsa. In *ECCV*, 2006.

[4] X. Chen and G. J. Zelinsky. Real-world visual search is dominated by top-down guidance. *Vision Research*, 46:4118–4133, 2006.

[5] G. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. In *ICCV*, 2003.

[6] M. Everingham, L. V. Gool, C. K. I.Williams, J.Winn, and A. Zisserman. The pascal visual object classes challenge 2007 results.

[7] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[8] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254–1259, Nov 1998.

[9] T. Jost, N. Ouerhani, R. von Wartburg, R. Mri, and H. Hgli. Assessing the contribution of color in visual attention. *CVIU*, 100(1–2):107–123, 2005.

[10] F. S. Khan, J. van de Weijer, and M. Vanrell. Fusing vocabularies for texture categorisation. In *CVCRD*, 2008.

[11] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *PAMI*, 31(7):1294–1309, 2009.

[12] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *PAMI*, 27(8):1265–1278, 2005.

[13] D. G. Lowe. Distinctive image features from scale-invariant points. *IJCV*, 60(2):91–110, 2004.

[14] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, 2008.

[15] M. Marszalek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representation for visual object class recognition 2007. In *Visual recognition Challenge Workshop in conjuncture with ICCV*, 2007.

[16] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.

[17] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, , and L. V. Gool. A comparison of affine region detectors. *IJCV*, 65(1–2):43–72, 2005.

[18] M.-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *CVPR*, 2006.

[19] M.-E. Nilsback and A. Zisserman. Delving into the whorl of flower segmentation. In *BMVC*, 2007.

[20] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

[21] F. Perronnin. Universal and adapted vocabularies for generic visual categorization. *PAMI*, 30(7):1243–1256, 2008.

[22] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modelling scenes with local descriptors and latent aspects. In *ICCV*, 2005.

[23] P. Quelhas and J.-M. Odobez. Natural scene image modeling using color and texture visterms. In *CIVR*, 2006.

[24] J. K. Regan, R. A. Rensink, and J. J. Clark. Change-blindness as a result of mudsplashes. *Nature*, 34:398, 1999.

[25] I. Spence, P. Wong, M. Rusan, and N. Rastegar. How color enhances visual memory for natural scenes. *Psychological Science.*, 17:1–6, 2006.

[26] J. Tsotsos. Analyzing vision at the complexity level. *Behav. Brain Sci.*, 13:423–469, 1990.

[27] K. van de Sande, T. Gevers, and C. Snoek. Evaluation of color descriptors for object and scene recognition. In *CVPR*, 2008.

[28] J. van de Weijer and C. Schmid. Coloring local feature extraction. In *ECCV*, 2006.

[29] J. van de Weijer and C. Schmid. Applying color names to image description. In *ICIP*, 2007.

[30] J. van de Weijer, C. Schmid, and J. J. Verbeek. Learning color names from real-world images. In *CVPR*, 2007.

[31] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *IJCV*, 72(2):133–157, 2007.

[32] F. A. Wichmann, L. T. Sharpe, and K. R. Gegenfurtner. The contributions of color to recognition memory for natural scenes. *Journal of Experimental Psychology:Learning, Memory, and Cognition.*, 28:509–520, 2002.

[33] J. M. Wolfe. *Visual Search*. 1998. in Attention, edited by H. Pashler, Psychology Press Ltd.

[34] J. M. Wolfe. *The Deployment of Visual Attention:Two Surprises*. Search and Target Acquisition, edited by NATO-RTO, NATO-RTO., 2000.

[35] J. M. Wolfe and T. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5:1–7, 2004.