Noname manuscript No.
(will be inserted by the editor)

Modulating Shape Features by Color Attention for Object Recognition

Fahad Shahbaz Khan, Joost van de Weijer, Maria Vanrell

Received:

Abstract Bag-of-words based image representation is a successful approach for object recognition. Generally, 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, it was found that the combination of different image cues, such as shape and color, often obtains below expected results.

This paper presents a novel method for recognizing object categories when using multiple cues by separately processing the shape and color cues and combining them by modulating the shape features by category-specific color attention. Color is used to compute bottom-up and top-down attention maps. Subsequently, these color attention maps are used to modulate the weights of the shape features. In regions with higher attention shape features are given more weight than in regions with low attention.

We compare our approach with existing methods that combine color and shape cues on five data sets containing varied importance of both cues, namely, Soccer (color predominance), Flower (color and shape parity), PASCAL VOC 2007 and 2009 (shape predominance) and Caltech-101 (color co-interference). The experiments clearly demonstrate that in all five data sets our proposed framework significantly outperforms existing methods for combining color and shape information.

Keywords Color Features, Image Representation, Object Recognition.

1 Introduction

Object category recognition is one of the fundamental problems in computer vision. In recent years several effective techniques for recognizing object categories from real-world images have been proposed. The bag-of-features framework, where images are represented by a histogram over visual words, is currently one of the most successful approaches to object and scene recognition. Many features such as color, texture, shape, and motion have been used to describe visual information for object recognition. Within the bag-of-words framework the optimal fusion of multiple cues, such as shape, texture and color, still remains an active research domain (Burghouts and Geusebroek, 2009; Gehler and Nowozin, 2009; van de Sande et al, 2010). Therefore in this paper, we analyze the problem of object recognition within the bag-of-words framework using multiple cues, in particular, combining shape and color information.

There exist two main approaches to incorporate color information within the bag-of-words framework (Quelhas and Odobez, 2006; Snoek et al, 2005). The first approach called, *early fusion*, fuses color and shape at the feature level as a result of which a joint color-shape vocabulary is produced. The second approach, called *late fusion*, concatenates histogram representation of both color and shape, obtained independently. Early fusion provides a more discriminative visual vocabulary, but might deteriorate for classes which vary significantly over one of the visual cues. For example, man-made categories such as cars and chairs vary considerably in color. On the other hand, late fusion is expected to per-

Fahad Shahbaz Khan, Joost van de Weijer, Maria Vanrell
Computer Vision Centre Barcelona, Universitat Autonoma de Barcelona
Tel.: +34-93-5814095
E-mail: fahad,joost,maria@cvc.uab.es

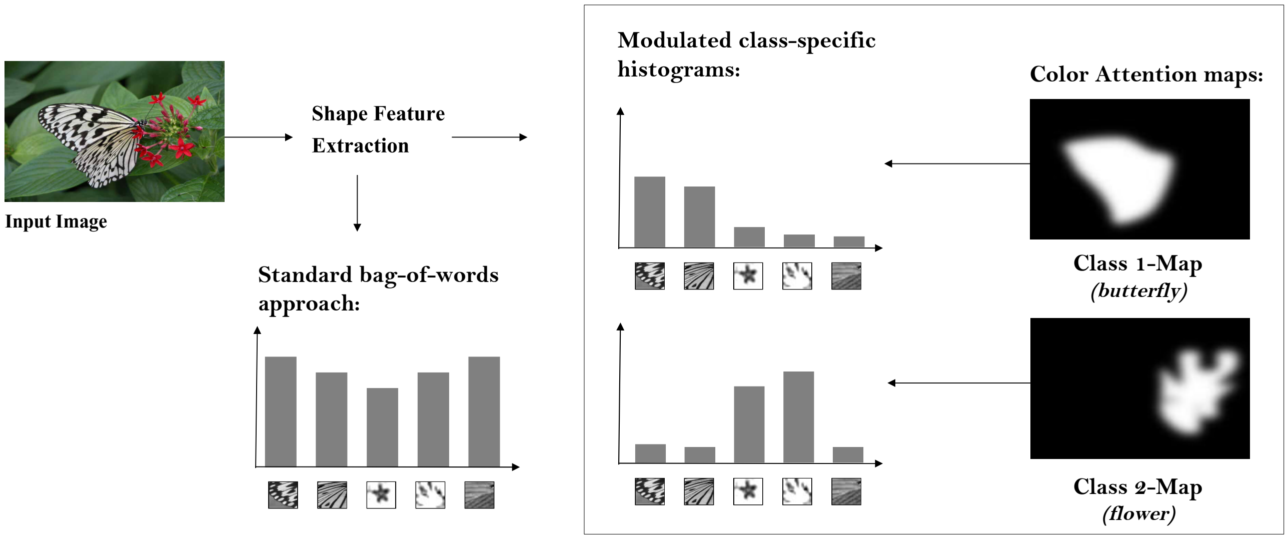

Fig. 1 Top-down control of visual attention based on color. In standard bag-of-words the image representation, here as distribution over visual shape words, is constructed in a bottom-up fashion. In our approach we use top-down class-specific color attention to modulate the impact of the shape-words in the image on 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.

form better for such classes, since it provides a more compact representation of both color and shape as separate visual vocabularies are constructed for individual cues. This prevents the different cues from getting diluted, which happens in case of a combined shape-color vocabulary. However, for classes which are characterized by both cues the visual vocabulary of late fusion will not be optimal. Such classes include natural categories like cats and trees which are better represented by early fusion based schemes.

Combining color and shape within the bag-of-words, using an early fusion based approach, has recently shown to provide excellent results on standard object recognition data sets (Everingham et al, 2008; van de Sande et al, 2010). Bosch et al (2008) propose to compute the SIFT descriptor in the HSV color space and concatenate the results into one combined color-shape descriptor. Photometrically invariant histograms are combined with SIFT for image classification by van de Weijer and Schmid (2006). A study into the photometric properties of many color descriptors and an extensive performance evaluation is performed by van de Sande et al (2008, 2010). In summary, most successful approaches (Bosch et al, 2008; van de Sande et al, 2010; van de Weijer and Schmid, 2006) proposed to combine color and shape features are based on early fusion scheme. As discussed before these early fusion methods are all expected to be suboptimal for classes where one of the cues varies significantly, like in the case of man-made objects.

This observation inspires us to propose a new image representation which combines multiple features within the bag-of-words framework. Our approach, *modulating shape features by color attention*, processes color and shape separately and combines them by means of bottom-up and top-down modulation of attention [1] as shown in Fig. 1. The top-down information is introduced by using learned class-specific color information to construct category-specific color attention maps of the categories. In Fig. 1 two color attention maps are visualized for the butterflies and flowers categories. Subsequently, this top-down color attention maps are used to modulate the weights of the bottom-up shape features. In regions with higher attention shape features are given more weight than in regions with low attention. As a result a class-specific image histogram is constructed for each category. We shall analyze the theoretical implications of our method and compare it to early and late fusion schemes used for combining color and shape features. Experiments will be conducted on standard object recognition data sets to evaluate the performance of our proposed method.

The paper is organized as follows. In Section 2 we discuss related work. In Section 3 the two existing approaches namely, early and late fusion, are discussed. Our approach is outlined based on an analysis of the relative merits of early and late fusion techniques in

[1] Throughout this paper we consider information which is dependent on the category-label as top-down, and information which is not as bottom-up.

Section 4. Section 5 starts with an introduction to our experimental setup followed by data sets used for our experiments and finally experimental results are given. Section 6 finishes with concluding remarks.

2 Related Work

There has been a large amount of success in using the bag-of-visual-words framework for object and scene classification (Bosch et al, 2006; Dorko and Schmid, 2003; Fei-Fei and Perona, 2005; Lazebnik et al, 2005; Mikolajczyk and Schmid, 2005; Quelhas et al, 2005; van de Weijer and Schmid, 2006) 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 vocabulary. 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.

Initially, many methods only used the shape features, predominantly SIFT (Lowe, 2004) to represent an image (Dorko and Schmid, 2003; Fei-Fei and Perona, 2005; Lazebnik et al, 2005). However, more recently the possibility of adding color information has been investigated (Bosch et al, 2006; Burghouts and Geusebroek, 2009; van de Sande et al, 2010; van de Weijer and Schmid, 2006). Previously, both early and late fusion schemes have been evaluated for image classification (Quelhas and Odobez, 2006). The comparison performed in recent studies suggest that combining multiple cues usually improves final classification results. However, within the bag-of-words framework the optimal fusion of different cues, such as shape, texture and color, still remains open to debate.

Several approaches have been proposed recently to combine multiple features at the kernel level. Among these approaches, multiple kernel learning, MKL, is the most well-known approach and significant amount of research has been done to exploit kernel combinations carrying different visual features (Bach, 2008; Bosch et al, 2007b; Rakotomamonjy et al, 2007; Varma and Babu, 2009; Varma and Ray, 2007). Other than MKL, averaging and multiplying are the two straight-forward and earliest approaches to combine different kernel responses in a deterministic way. Surprisingly, in a recent study performed by Gehler and Nowozin (2009) it has been shown that in some cases the product of different kernel responses provide similar or even better results than MKL. It is noteworthy to mention that our approach is essentially different from MKL because it proposes a new image representation. Like early and late fusion it can further be used as an input to an MKL.

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 (2009) propose to learn discriminative visual vocabularies, which are optimized to separate the class labels. Perronnin (2008) 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 (2007) semantically label local features into a number of semantic concepts for the task of scene classification. Yang et al (2008) propose an optimization method to unify the visual vocabulary construction with classifier training phase. Fulkerson et al (2008) propose a method to generate compact visual vocabularies based on agglomerative information bottleneck principle. This method defines the discriminative power of a visual vocabulary as the mutual information between a visual word and a category label.

There have been several approaches proposed in recent years to learn an efficient visual codebook for image classification and retrieval tasks. Sivic and Zisserman (2003) propose an approach to object matching in videos by using inverted file system and document ranking. Winn et al (2005) propose a method for image categorization by learning appearance-based object models from training images. A large vocabulary is compressed into a compact visual vocabulary by learning a pairwise merging of visual-words. Jurie and Triggs (2005) argue that visual vocabulary based on standard k-means algorithm on densely sampled patches provides inferior performance and propose an acceptance-radius based clustering approach for recognition and detection. Tuytelaars and Schmid (2007) propose a data independent approach to construct a visual vocabulary by discritizing the feature space using a regular lattice for image classification. Cai et al (2010) propose an approach for estimating codebook weights especially in scenarios when there are insufficient training samples to construct a large size visual codebook. The above-mentioned approaches mainly aim at improving the visual codebook construction stage, whereas the novelty of our proposed method is that we use feature weighting as a mechanism to bind color and shape visual cues.

Humans have an outstanding ability to perform various kinds of visual search tasks constantly. But how is it that the human visual system does this job with little effort and can recognize a large number of object categories with such an apparent ease? Research on the human vision system suggests that basic visual features such as shape and color are processed in parallel, and are not combined in an early fusion manner. For example, in the two-stage architecture of the well known *Feature Integration Theory* by Treisman (1996), the processing of basic features in an initially parallel way is done in the first stage, also known as the "preattentive stage". These basic visual features processed separately are loosely bundled into objects before they are binded into a recognizable object (Wolfe, 2000; Wolfe and Horowitz, 2004). It is further asserted that the basic features are initially represented separately before they are integrated at a later stage in the presence of attention. Similarly, we propose a framework where color and shape are processed separately. Other than late fusion, where histograms of individual features are concatenated after processing, we propose to combine color and shape by separately processing both visual cues and then modulating the shape features using color as an attention cue.

Several computational models of visual attention have been proposed previously. The work of Tsotsos et al (1995) uses top-down attention and local winner-take-all networks for tuning model neurons at the attended locations. Itti et al (1998) propose a model for bottom-up selective visual attention. The visual attention mechanism has been based on serial scanning of a saliency map computed from local feature contrasts. The saliency map computed is a two-dimensional topographic representation of conspicuity or saliency for every pixel in the image. The work was further extended by Walther and Koch (2006) from salient location to salient region-based selection. Meur et al (2006) propose a coherent computational approach to the modeling of bottom-up visual attention where contrast sensitivity functions, perceptual decomposition, visual masking, and center-surround interactions are some of the features implemented in the model. Peters and Itti (2007) introduce a spatial attention model that can be applied to both static and dynamic image sequences with interactive tasks. Gao et al (2009) propose a top-down visual saliency framework that is intrinsically connected to the recognition problem and closely resembles to various classical principles for the organization of perceptual systems. The method aims at two fundamental problems in discriminant saliency, feature selection and saliency detection. In summary, the visual attention phenomenon has been well studied in the fields of psychology and neuroscience but still has not been investigated within the bag-of-words framework for combining multiple visual cues.

This paper is an extended version of our earlier work (Khan et al, 2009). We extended the model by introducing a bottom-up component of attention. In our new model both bottom-up and top-down components of color attention are employed to modulate the weights of local shape features. Moreover, we introduce two parameters to tune the relative contribution of the two attention components. The first parameter controls the influence of color and shape information. The second parameter is employed to leverage the contribution of top-down and bottom-up attention mechanisms. Finally, we have extended the experiments with results on the Caltech-101 data set.

3 Early and Late Feature Fusion

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.

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, where M^i is the total number of features in image i. Examples of commonly used detectors are multi-scale grid sampling and interest point detectors such as Laplace and Harris corner detector. Generally, the local features are represented in visual vocabularies which describe various image cues such as shape, texture, and color. We 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, $\mathrm{W}^\mathrm{k} = \{\mathrm{w}_1^\mathrm{k}, ..., \mathrm{w}_{\mathrm{V}^\mathrm{k}}^\mathrm{k}\}$, with the visual words $\mathrm{w}_\mathrm{n}^\mathrm{k}$, n=1,2,...,$V^k$ and $k \in \{s, c, sc\}$ for the two separate cues shape and color and for the combined visual vocabulary of color and shape. The local features f_{ij} are quantized differently for the two approaches: by a pair of visual words $(\mathrm{w}_\mathrm{ij}^\mathrm{s}, \mathrm{w}_\mathrm{ij}^\mathrm{c})$ for late fusion and by single shape-color word $\mathrm{w}_\mathrm{ij}^\mathrm{sc}$ in the case of early fusion. Thus, $\mathrm{w_{ij}}^\mathrm{k} \in \mathrm{W}^\mathrm{k}$ is the j^{th} quantized feature of the i^{th} image for a visual cue k.

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

$$h\left(\mathrm{w}_n^\mathrm{k}|I^i\right) \propto \sum_{j=1}^{M^i} \delta\left(\mathrm{w}_\mathrm{ij}^\mathrm{k}, \mathrm{w}_\mathrm{n}^\mathrm{k}\right) \tag{1}$$

with

$$\delta\left(x, y\right) = \begin{cases} 0 & \text{for } x \neq y \\ 1 & \text{for } x = y \end{cases} \tag{2}$$

For early fusion, thus called because the cues are combined before vocabulary construction, we compute histogram $h\left(\mathrm{w}^\mathrm{sc}|I^i\right)$. For late fusion we compute histograms

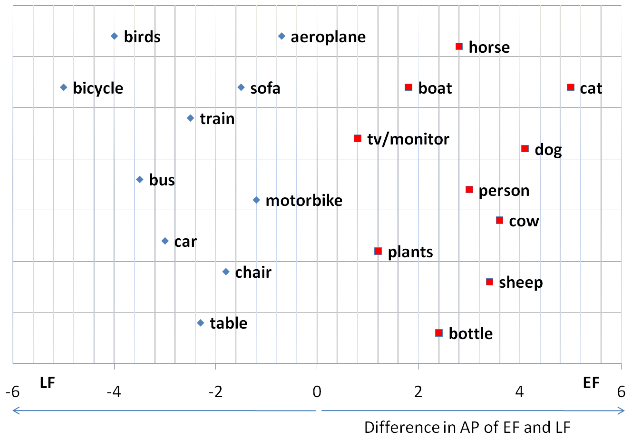

Fig. 2 Difference in average precision (AP) scores of early and late fusion schemes for the 20 categories of PASCAL VOC 2007 data set. Vertical axis does not contain information. Half of the categories are better represented by early fusion (red) and half by late fusion(blue).

$h\left(\mathrm{w}^{\mathrm{s}}|I^i\right)$ and $h\left(\mathrm{w}^{\mathrm{c}}|I^i\right)$ and concatenate the distributions. It is important to introduce a parameter balancing the relative weight between the different cues. For the results of early and late fusion reported in this paper we learn this parameter by means of cross-validation on the validation set.

Late and early fusion methods lead to different image representations and therefore favor different object categories. To better understand their strengths we perform an experiment on the PASCAL VOC 2007 data set which contains a wide variety of categories. Both early and late fusion results are obtained using SIFT and Color Names descriptors. The results are presented in Fig 2. The axis shows the difference between the average precision (AP) scores of early and late fusion schemes (e.g. bicycle has a 5% higher score when represented by late fusion than by early fusion, and for airplane both representation yield similar results). The results clearly show that neither of the two fusion approaches perform well for all object categories.

Most man-made categories namely, bicycle, train, car and buses performs better with late fusion over its early fusion counterpart. The only exception in this case is the boat category which is better represented by early fusion. On the other hand, natural categories such as cow, sheep, dog, cat, horse etc. are better represented by early fusion. The bird category is the only outlier among natural categories which provides superior performance with late fusion instead of early fusion. Better than the distinction between man-made and natural categories is the distinction between color-shape dependency and color-shape independency of categories. This explains the location of most of the categories along the axes, including the birds class which is represented by a large variety of bird species with widely divergent colors, and the boat class which contains mainly white boats. The difference in strength of early and late fusion on different object categories is illustrated in Fig 3.

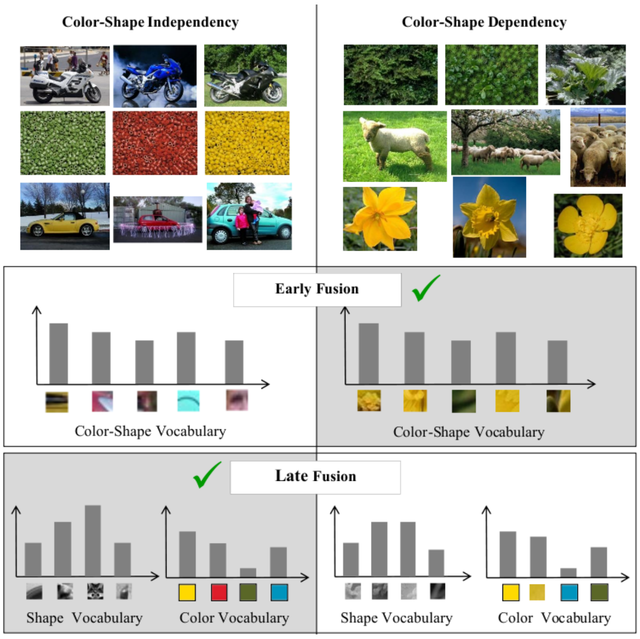

Fig. 3 Graphical explanation of early and late fusion approaches. Note that for some classes early fusion scheme performs better where as for some categories, late fusion outperforms early fusion methods.

Based on the above analysis of early and late fusion we conclude that, to combine multiple cues, two properties are especially desired. The first property is *feature compactness*. Having this property implies constructing a separate visual vocabulary for both color and shape. This is especially important for classes which have color-shape independency. Learning these classes from a combined shape-color vocabulary only complicates the task of the classifier. Late fusion possesses the property of feature 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 (as desired for categories with color-shape dependency). 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.

4 Color Attention for Object Recognition

In the previous section we elaborated two approaches to combine color and shape features. In this section, we propose an attention-based image representation. Feature binding and feature compactness will be achieved by modulating shape features with bottom-up and top-down components of color attention.

4.1 Attention-based Bag-of-Words

We define a generalization of the bag-of-words as given by Eq. 3, called *attention-based bag-of-words*:

$$h\left(\mathrm{w}_n^{\mathrm{k}}|I^i\right) \propto \sum_{j=1}^{M^i} a_{ij}\delta\left(\mathrm{w}_{\mathrm{ij}}^{\mathrm{k}}, \mathrm{w}_{\mathrm{n}}^{\mathrm{k}}\right), \qquad (3)$$

where a_{ij} are the attention-weights which modulate feature $\mathrm{w}_{\mathrm{ij}}^{\mathrm{k}}$. Choosing the a_{ij} weights to be equal to one reduces the equation to standard bag-of-words. The weights can be interpreted as attention maps, essentially determining which features w^{k} are relevant.

Next, we apply attention-based bag-of-words to combine color and shape. For this purpose we separate the functionality of the two visual cues. The shape cue will function as *descriptor cue*, and is used similar as in the traditional bag-of-words. The color cue is used as an *attention cue*, and determines the impact of the local features on the image representation. To obtain our image representation, color attention is used to modulate the shape features according to:

$$h\left(\mathrm{w}_n^{s}|I^i, class\right) \propto \sum_{j=1}^{M^i} a\left(\mathbf{x}_{ij}, class\right)\delta\left(\mathrm{w}_{\mathrm{ij}}^{\mathrm{s}}, \mathrm{w}_{\mathrm{n}}^{\mathrm{s}}\right), \qquad (4)$$

where $a\left(\mathbf{x}_{ij}, class\right)$ denotes the color attention of the j^{th} local feature of the i^{th} image and is dependent on both the location $\mathbf{x}_{ij}$ and the *class*. The difference to standard bag-of-words is that in regions with high attention, shape-features are given more weight than in regions with low attention. This is illustrated in the two attention-based bag-of-words histograms in Fig. 1 where the attention map of the butterfly results in a bag-of-words representation with an increased count for the visual words relevant to butterfly (and similarly for the flower representation). Note that all histograms are based on the same set of detected shape features and only the weighting varies for each *class*. As a consequence a different distribution over the same shape words is obtained for each *class*.

Similarly as for human vision we distinguish between bottom-up and top-down attention:

$$a\left(\mathbf{x}_{ij}, class\right) = a_b\left(\mathbf{x}_{ij}\right) a_t\left(\mathbf{x}_{ij}, class\right). \qquad (5)$$

Here $a_b\left(\mathbf{x}_{ij}\right)$ is the bottom-up color attention based on the image statistics and highlights the most salient color locations in an image. The top-down color attention is represented by $a_t\left(\mathbf{x}_{ij}, class\right)$, describing our prior knowledge about the color appearance of the categories we are looking for. The two components will be discussed in detail later.

Two parameters are introduced to tune the relative contribution of the two attention components:

$$a\left(\mathbf{x}_{ij}, class\right) = \left(a_b\left(\mathbf{x}_{ij}\right)^{(1-\beta)} a_t\left(\mathbf{x}_{ij}, class\right)^{\beta}\right)^{\gamma}. \qquad (6)$$

The parameter, γ, is used to control the influence of color versus shape information. For $\gamma = 0$ we obtain a standard bag-of-words based image representation where a higher value of γ denotes more influence of color attention. The second parameter, β, is employed to vary the contribution of top-down and bottom-up attention, where $\beta = 0$ indicates only bottom-up attention and $\beta = 1$ means only top-down attention. Both γ and β parameters are learned through cross-validation over the validation set.

The image representation proposed in Eq. 4 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, our image representation does combine the two properties *feature binding* and *feature compactness*. Firstly, *feature compactness* is achieved since we construct separate visual vocabularies for both color and shape cues. Secondly, *feature binding* is achieved by the top-down modulation as follows from Eq. 4. Consequently, we expect to obtain better results by combining both these properties into a single image representation.

The attention framework as presented in Eq. 3 recalls earlier work on the feature weighting techniques (Wettschereck et al, 1997). Replacing $a_{ij} = a_n$ transforms the equation to a classical feature weighting scheme in which separate weights for each feature are introduced, allowing to leverage their relative importance and reduce the impact of noisy features. The main difference with our approach is twofold. Firstly, our weighting is dependent on the position in the image (as indexed by i) which allows for the feature binding. Secondly, we use a different cue, the attention cue, to compute the weight. As a consequence, the final image representation is based on the combination of the two cues, color and shape.

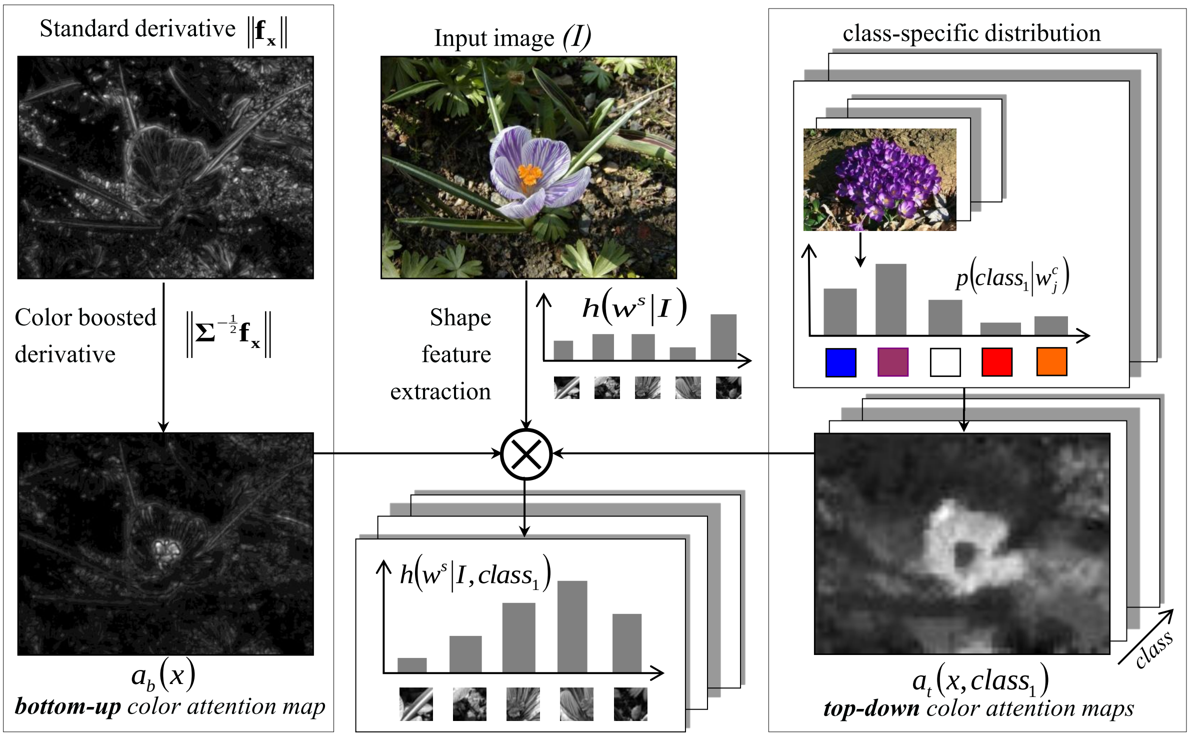
Standard derivative $\|\mathbf{f}_{\mathbf{x}}\|$
Input image *(I)*
class-specific distribution
Color boosted derivative
$\left\|\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{f}_{\mathbf{x}}\right\|$
Shape feature extraction
$h\left(w^{s}|I\right)$
$p\left(class_1|w_j^c\right)$
$h\left(w^{s}|I, class_1\right)$
$a_b(x)$
***bottom-up** color attention map*
$a_t(x, class_1)$
***top-down** color attention maps*
class

Fig. 4 An overview of our method. Other than the classical bag-of-words approach, our method modulates the shape features with bottom-up and top-down color attention. Bottom-up attention is based on image statistics to indicate the most salient color regions whereas the top-down attention maps provide class-specific color information. As a result, a class-specific histogram is constructed by giving prominence to those shape visual-words that are considered relevant by the attention maps.

4.2 Top-down Color Attention

Here we define the top-down component of color attention of local features to be equal to the probability of a class given its color values and it is defined by:

$$a_t\left(\mathbf{x}_{ij}, class\right) = p\left(class|\mathrm{w}^{\mathrm{c}}_{\mathrm{ij}}\right). \tag{7}$$

The local color features at the locations $\mathbf{x}_{ij}$ are vector quantized into a visual vocabulary where $\mathrm{w}^{\mathrm{c}}_{\mathrm{ij}}$ describes a visual word. The probabilities $p\left(class|\mathrm{w}^{\mathrm{c}}_{\mathrm{ij}}\right)$ are computed using Bayes theorem,

$$p\left(class|\mathrm{w}^{\mathrm{c}}\right) \propto p\left(\mathrm{w}^{\mathrm{c}}|class\right) p\left(class\right) \tag{8}$$

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

$$p\left(\mathrm{w}^{\mathrm{c}}_{\mathrm{n}}|class\right) \propto \sum_{I^{class}} \sum_{j=1}^{M^{i}} \delta\left(w^{k}_{ij}, \mathrm{w}^{\mathrm{c}}_{\mathrm{n}}\right), \tag{9}$$

obtained by summing over the indexes of the training images for the category I^{class} . The prior over the classes $p\left(class\right)$ is obtained from the training data. For categories where color is irrelevant, $p\left(class|\mathrm{w}^{\mathrm{c}}\right)$ is uniform and our model simplifies to the standard bag-of-words representation. If the bounding box information is available it was found that the probabilities computed only from features inside the bounding boxes provide better results. Thus when available we used bounding box knowledge available to obtain the probabilities.

If we compute $p\left(class|\mathrm{w}^{\mathrm{c}}\right)$ for all local features in an image we can construct a top-down class-specific color attention map. Several examples are given in Fig. 5. The color attention map is used to modulate the local shape features. Each category provides its own attention map, consequently, a different histogram is constructed for each category. The final image representation is constructed by concatenating the category-specific histograms. The image representation is normalized before classification.

4.3 Bottom-up Color Attention

Bottom-up attention is employed to determine salient locations obtained from visual features such as color, intensity, orientation etc in an image. Contrary to top-down attention, bottom-up attention is independent of the object categories since it is not task dependent. In this work, we apply the color saliency boosting method (van de Weijer et al, 2006) to compute bottom-up attention maps. The color saliency boosting algorithm is based on the application of information theory to the

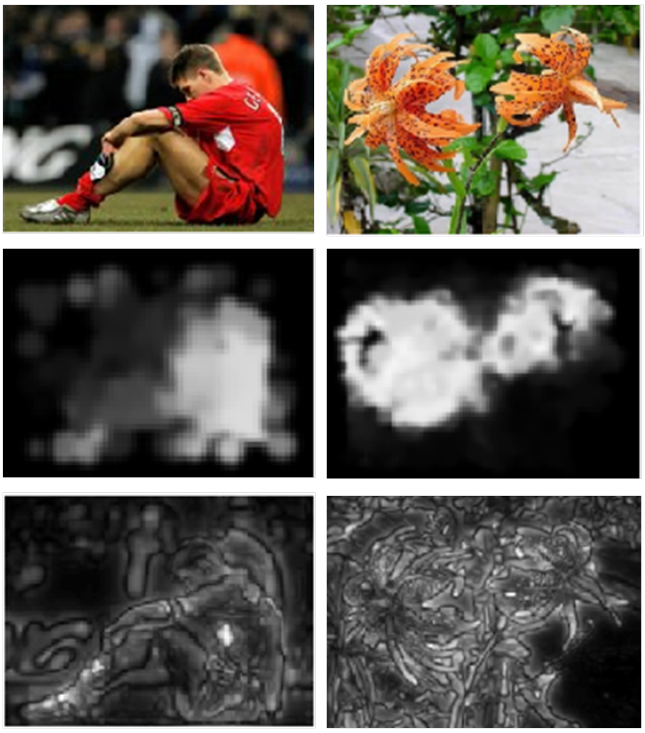

Fig. 5 Top-down color attention and bottom-up saliency maps. First row: a Liverpool class category image from soccer data set and a Tiger lily flower species image from flower data set. Second row: Top-down color attention maps of the images. Third row: Bottom-up saliency map of the images.

statistics of color image derivatives. It has been successfully applied to image retrieval and image classification (Stottinger et al, 2009; van de Sande et al, 2010).

Let $\mathbf{f_x} = (R_\mathbf{x}\ G_\mathbf{x}\ B_\mathbf{x})^T$ be the spatial image derivatives. The information content of first order derivatives in a local neighborhood is given by

$$I(\mathbf{f_x}) = -log(p(\mathbf{f_x})) \tag{10}$$

where $p(\mathbf{f_x})$ is the probability of the spatial derivative. The equation states that a derivative has a higher information content if it has a low probability of occurrence. In general, the statistics of color image derivatives are described by a distribution which is dominated by a principal axis of maximum variation along the luminance direction, and two minor axes, attributed to chromatic changes. This means that changes in intensity are more probable than chromatic changes and therefore contain less information content. The color derivative distribution can be characterized by its second-order statistics, i.e. its covariance matrix $\mathbf{\Sigma_x} = E[\mathbf{f_x f_x^T}]$. When we apply a whitening transformation to the image derivatives according to, $\mathbf{g_x} = \mathbf{\Sigma_x}^{-\frac{1}{2}}\mathbf{f_x}$, this will result in a more homogeneous derivative distribution for $\mathbf{g_x}$, in which the dominant variations in the intensity axes are suppressed, and the chromatic variations are enforced. As a result points with equal derivative strength, $\|\mathbf{g_x}\|$, have similar information content.

Similar as in Vazquez et al (2010) we apply color boosting to compute a multi-scale contrast color attention map:

$$a_b(\mathbf{x}) = \sum_{\sigma\in S}\sum_{\mathbf{x}'\in N(\mathbf{x})} \left\| (\mathbf{\Sigma}_x^\sigma)^{-\frac{1}{2}} \left(\mathbf{f}^\sigma(\mathbf{x}) - \mathbf{f}^\sigma(\mathbf{x}')\right) \right\| \tag{11}$$

where $\mathbf{f}^\sigma$ is the Gaussian smoothed image at scale σ, $N(\mathbf{x})$ is a 9x9 neighborhood window, moreover $S = \left[1, \sqrt{2}, 2, 2\sqrt{2}, ..., 32\right]$. We compute $\mathbf{\Sigma_x^\sigma}$ from the derivatives at scale σ from a single image. The approach is an extension of the multi-contrast method by Liu et al (2007) to color. Examples of bottom-up attention maps are given in Fig. 5. These images demonstrate that the dominant colors are suppressed and the colorful, less frequent, edges are enhanced.

4.4 Multiple Cues

The proposed method can easily be extended to include multiple bottom-up and top-down attention cues. In this paper we have also evaluated multiple top-down attention cues. For q top-down attention cues we compute

$$a\left(\mathbf{x}_{ij}, class\right) = a_t^1\left(\mathbf{x}_{ij}, class\right) \times ... \times a_t^q\left(\mathbf{x}_{ij}, class\right). \tag{12}$$

Note that the dimensionality of the image representation is independent of the number of attention cues. In the experiments, we shall provide results based on multiple color attention cues.

4.5 Relation to Interest Point Detectors

In bag-of-words two main approaches to feature detection can be distinguished (Mikolajczyk et al, 2005). Ignoring the image content *dense sampling* extracts features on a dense grid at multiple scales in the image. *Interest point* detectors adjust to the image by sampling more points from regions which are expected to be more informative. Examples of the most used interest point detectors are Harris-Laplace, Hessian and Laplace detectors. Here we show that interest point detectors can also be interpreted to be a shape-attention weighted version of a dense multi-scale feature detector.

Consider the following equation for attention based bag-of-words:

$$h\left(\mathrm{w}_n^s | I^i, class\right) \propto \sum_{j=1}^{M^i} a\left(\mathbf{x}_{ij\sigma}\right) \delta\left(\mathrm{w^s_{ij\sigma}}, \mathrm{w^s_n}\right), \tag{13}$$

where σ has been added to explicitly indicate that at every location multiple scales are taken into consideration. Interest point detectors can be considered as providing the function $a\left(\mathbf{x}_{ij\sigma}\right)$ which is one for feature locations and scales which were detected and zero otherwise. For example the Laplace detector computes the function $a\left(\mathbf{x}_{ij\sigma}\right)$ by finding the maxima in the Laplace scale-space representation of the image, and thereby providing a scale invariant blob detector. In these cases the shape-attention is bottom-up since the same detector is used invariably for all classes. The importance of interest point detectors versus dense sampling is much researched (Mikolajczyk et al, 2005; Marszalek et al, 2007; Nowak et al, 2006) and is not further investigated in this paper.

Of interest here is the insight this gives us in the working of color attention. Although color attention does not have the hard assignment which is applied in traditional interest point detectors (selecting some features and ignoring others), the weights $a\left(\mathbf{x}_{ij}, class\right)$ could be understood as a color based 'soft' interest point detector, where some features have more weights than others. Furthermore, since the weights are class dependent, the resulting histograms can be interpreted as being formed by class-specific interest point detectors.

5 Experiments

In this section we first explain the experimental setup followed by an introduction to the data sets used in our experiments. The data sets have been selected to represent a varied importance of the two visual cues namely, color and shape. We then present the results of our proposed method on image classification. Finally, the results are compared to state-of-the-art methods fusing color and shape.

5.1 Experimental Setup

To test our method, we have used a standard multiscale grid detector along with Harris-Laplace point detector (Mikolajczyk et al, 2005) and a blob detector. We normalized all the patches to a standard size and descriptors are computed for all regions in the feature description step. A universal visual vocabulary representing all object categories in a data set is then computed by clustering the descriptor points using a standard K-means algorithm. In our approach the SIFT descriptor is used to create a shape vocabulary. A visual vocabulary of 400 is constructed for Soccer and Flower data sets. For Pascal VOC 2007 and 2009 data sets, a 4000 visual-word vocabulary is used. A visual vocabulary of 500 is employed for the Caltech-101 data set. To construct a color vocabulary, two different color descriptors, namely the color name (CN) descriptor (van de Weijer and Schmid, 2007; van de Weijer et al, 2009) and hue descriptor (HUE) (van de Weijer and Schmid, 2006). Since color names has more discriminative power than hue we used a larger vocabulary for CN than for HUE for all datasets.

We shall abbreviate our results with the notation convention $CA(descriptor\ cue, attention\ cues)$ where CA stands for the integrated bottom-up and top-down components of color attention based bag-of-words and $TD(descriptor\ cue, attention\ cue)$ where TD stands for Top-Down attention based bag-of-words representation. We shall provide results with one attention cue $CA(SIFT,\ HUE)$, $CA(SIFT,\ CN)$, and color attention with two attention cues $CA(SIFT, \{HUE, CN\})$ combined by using Eq. 12. The final image representation input to an SVM classifier is equal to the size of shape vocabulary times the number of object categories in the data set. In our experiments we use a standard non-linear SVM. A single γ and β parameter is learned for Soccer and Flower data set. For Caltech-101 parameters are learned globally for the whole data set whereas for the PASCAL VOC data sets class-specific parameters are learned.

We compare our method with the standard methods used to combine color and shape features from literature: early fusion and late fusion. We perform early and late fusion with both CN and HUE descriptors. We also compare our approach with methods that combine color and shape at the classification stage by combining the multiple kernel responses. Recently, an extensive performance evaluation of color descriptors has been presented by van de Sande et al (2010). We 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 C-SIFT 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 by van de Sande et al (2010) that with no prior knowledge about object categories, OpponentSIFT descriptor was found to be the best choice.

5.2 Image Data Sets

We tested our method on five different and challenging data sets namely Soccer, Flower, PASCAL VOC 2007 and 2009 and Caltech-101 data sets. The data sets vary

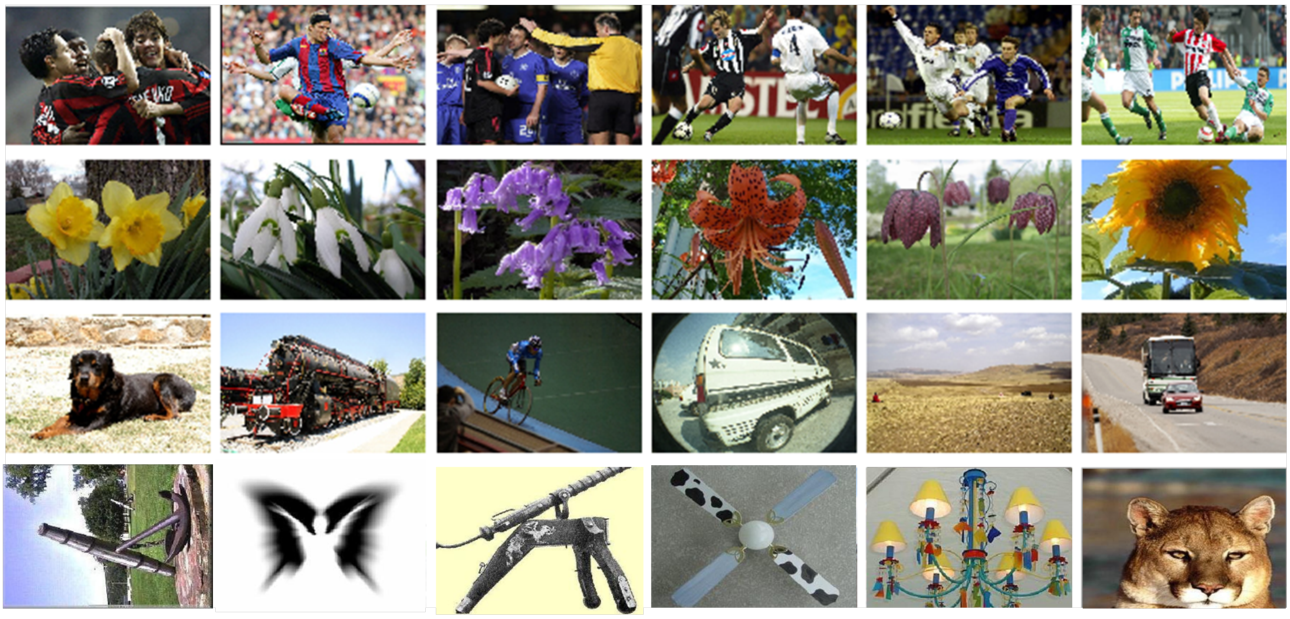

Fig. 6 Examples from the four data sets. From top to bottom: Soccer, Flower, PASCAL VOC and Caltech-101 data sets.

in the relative importance of the two cues, shape and color.

The Soccer data set [2] consists of 7 classes of different soccer teams (van de Weijer and Schmid, 2006). Each class contains 40 images divided in 25 train and 15 test images per class. The Flower data set [3] consists of 17 classes of different variety of flower species and each class has 80 images. We use both the 40 training and 20 validation images per class (60) to train (Nilsback and Zisserman, 2006). We also tested our approach on PASCAL VOC data sets (Everingham et al, 2007, 2009). The PASCAL VOC 2007 data set [4] consists of 9963 images of 20 different classes with 5011 training images and 4952 test images. The PASCAL VOC 2009 data set [5] consists of 13704 images of 20 different classes with 7054 training images and 6650 test images. Finally, we tested our approach on Caltech-101 data set. The Caltech-101 data set [6] contains 9144 images of 102 different categories. The number of images per category varies from 31 to 800. Fig. 6 shows some images from the four data sets.

5.3 Attention Cue Evaluation

In this paper, we propose to combine color and shape by modulating shape features using color as an attention cue. The same framework can be used to modulate color features by exchanging the roles of color and shape. Table 1 provides results of our experiments where we investigate shape-shape attention, color-color attention, shape-color attention and color-shape attention. Experiments are performed on both Soccer and Flower data sets. The results in Table 1 suggest that color is the best choice as an attention cue, which coincides with the previous works done in visual attention literature (Wolfe and Horowitz, 2004; Jost et al, 2005). Therefore, in the following experiments color is used as an attention cue to modulate the shape features. [7]

$Attention - Cue$	$Descriptor - Cue$	$Soccer$	$Flower$
$Shape$	$Shape$	50	69
$Color$	$Color$	79	66
$Shape$	$Color$	78	69
$Color$	$Shape$	**87**	**87**

Table 1 Classification Score (percentage) on Soccer and Flower Set Data sets. The results are based on top-down color attention obtained by using different combinations of color and shape as attention and descriptor cues.

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

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

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

[5] The PASCAL VOC Challenge 2009 at http://www.pascal-network.org/challenges/VOC/voc2009/

[6] The Caltech-101 object category data set at http://www.vision.caltech.edu/ImageDatasets/Caltech101/

[7] In an additional experiment, we tried improving the results by using a color-shape descriptor cue and an attention cue. This was found to deteriorate the recognition performance.

5.4 Soccer Data Set: 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 2. The importance of color for this data set is demonstrated by the unsatisfactory results of shape alone where an accuracy of 50% is obtained. 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. In all cases combining features by color attention was found to outperform both early and late fusion. We also combine color and shape by taking the product of the two kernels obtaining a classification score of 91%. Note that also for both early and late fusion the relative weight of color and shape features is learned by cross-validation. The best results are obtained by combining the top-down and bottom-up attention demonstrating the fact that both types of attentions are important for obtaining best classification results.

Our method outperforms the best results reported in literature (van de Weijer and Schmid, 2007), where a score of 89% is reported, based on a combination of SIFT and CN in an early fusion manner. Further we compare to C-SIFT and Opp-SIFT (van de Sande et al, 2010) which provide an accuracy of 72% and 82% respectively. The below expected results for C-SIFT might be caused by the importance of the achromatic colors to recognize the team shirts (for example, Milan outfits are red-black and PSV outfits are red-white). This information is removed by the photometric invariance of C-SIFT. Our best results of 96% is obtained when color has greater influence over shape (γ=3) which is also analogous to the unsatisfactory results of shape alone. Moreover, top-down attention has more influence than bottom-up attention (β=0.6).

Method	(SIFT,HUE)	(SIFT,CN)	(SIFT,(CN,HUE))
$EarlyFusion$	84	88	90
$LateFusion$	81	86	88
TD	87	90	94
CA	**90**	**91**	**96**

Table 2 Classification scores (percentage) for various fusion approaches on Soccer Data set. The best results are obtained by CA outperforming the other fusion methods by 5%.

5.5 Flower Data Set: color and shape parity

Image classification results on the Flower data set show the performance of our method on a data set for which both shape and color information are essential. The task is to classify the images into 17 different categories of flower-species. The use of both color and shape are important as some flowers are clearly distinguishable by shape, e.g. daisies and some other by color, e.g. fritillaries.

Method	(SIFT,HUE)	(SIFT,CN)	(SIFT,(CN,HUE))
$EarlyFusion$	87	88	89
$LateFusion$	86	87	88
TD	90	90	91
CA	**93**	**94**	**95**

Table 3 Classification Scores (percentage) for various fusion approaches on Flower Data set. CA is shown to outperform existing fusion approaches by 6%.

The results on flower data set are given in Table 3. As expected on this data set early fusion provides better results compared to late fusion. [8]. Again combining color and shape by color attention obtains significantly better results than both early and late fusion. We also significantly outperform both C-SIFT and OpponentSIFT which provide classification scores of 82% and 85% respectively.

On this data set our method surpassed the best results reported in literature (Nilsback and Zisserman, 2008; Xie et al, 2010; Orabona et al, 2010; Gehler and Nowozin, 2009). The results reported on this data set by Nilsback and Zisserman (2008) is 88.3% where shape, color and texture descriptors were combined along with the segmentation scheme proposed byNilsback and Zisserman (2007). 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 Xie et al (2010) where a result of 89.02% was obtained by combining the spatial pyramids of SIFT with OpponentSIFT, C-SIFT, rgSIFT and RGBSIFT respectively using a bin-ratio dissimilarity kernel. [9]

In Fig. 7 the classification score as a function of γ and β is provided. Our best result of 95% is obtained with a significant color influence (γ=2). Moreover, for

[8] We also performed an experiment for combining our color and shape features by using MKL. However, slightly better results of 86% were obtained by using a simple product of different kernel combinations which is similar to the results provided by Gehler and Nowozin (2009).

[9] The result reported by Ito and Kubota (2010) is not the recognition score commonly used to evaluate the classification performance on Flower data set and therefore is not compared with our approach in this paper.

this data set bottom-up attention has the same influence as top-down attention (β=0.5). It can also be seen that bottom-up attention alone improves results from 69% to 76%.

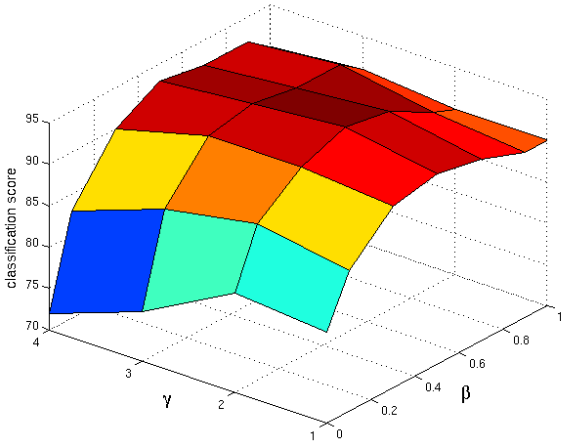

Fig. 7 Recognition performance as a function of γ and β for the Flower data set. From a shape only representation (γ=0 and β=0) the score goes up from 69% to 95% by leveraging the influence of color versus shape and the two components of color attention.

5.6 PASCAL VOC Data Sets: shape predominance

We test our approach where the shape cue is predominant and color plays a subordinate role and report image classification results on the PASCAL VOC 2007 and 2009 data sets. The PASCAL VOC 2007 data set contains nearly 10,000 images of 20 different object categories. The 2009 PASCAL VOC data set contains 13704 images of 20 different categories. For these data sets the average precision is used as a performance metric in order to determine the accuracy of recognition results.

On this data set, shape alone provides a MAP of 53.7 on this data set. A MAP of 49.6 is obtained using C-SIFT. This drop in performance is caused by the categories having color-shape independency which effects early fusion based approaches. Table 4 shows the results of different color-shape fusion schemes. Among the existing approaches late fusion provides the best recognition performance of 56.0. Our proposed framework obtains significantly better results and doubles the gain obtained by color. Our best results of 58.0 is obtained by the combination of bottom-up and top-down attention. For categories such as plants and tvmonitor, color is more important than shape (γ=3) where as for categories like sheep, sofa and cars shape is more influential as compared to color (γ=1). For categories such as cow, dogs and bottle bottom-up attention plays an important role. However, for most categories top-down attention plays a larger role than bottom-up attention on this data set.

Method	(SIFT,HUE)	(SIFT,CN)	(SIFT,(CN,HUE))
$EarlyFusion$	54.6	54.8	55.7
$LateFusion$	55.3	55.6	56.0
TD	56.6	56.8	57.5
CA	**57.0**	**57.5**	**58.0**

Table 4 Mean Average Precision on PASCAL VOC 2007 Data Set. Note that our results significantly improve the performance over the conventional methods of combining color and shape namely, Early and Late feature fusion.

The results per object category are given in Fig. 8. It is worthy to observe that our approach performs substantially better over early and late fusion approaches on a variety of categories. Recall that early fusion approaches lack feature compactness and struggle with categories where one cue is constant and the other cue varies considerably. This behavior can be observed in object categories such as motorbike, bird etc. In such classes early fusion provides below-expected results. On the other hand, late fusion lacks feature binding as it struggles over categories characterized by both color and shape. This is apparant in categories such as cat, sheep, cow where early fusion provides better results over late fusion. Our approach, which combines the advantages of both early and late fusion, obtains good results on most type of categories in this data set.

To illustrate the strength of different image representations, Table 5 shows images of different object categories from the PASCAL VOC 2007 data set. For this data set the average precision is used as an evaluation criteria. To obtain an average precision for each object category, the ranked output is used to compute the precision/recall curve. Table 5 shows example images from bird, pottedplant, sofa and motorbike categories and their corresponding ranks obtained from different methods. Early fusion performs better than late fusion on the pottedplant image since color remains constant (color-shape dependency). For the motorbike image, which possesses color-shape independency, late fusion performs best. Color attention outperforms other approaches on the first three example images.

The best entry in PASCAL 2007 VOC was by Marszalek et al (2007) where a mean average precision of 59.4 was reported by using SIFT, Hue-SIFT, spatial pyramid matching and a novel feature selection scheme.

Ranking of Different Object Categories

Method				
SIFT	1243	697	1325	155
Early Fusion	196	65	654	124
Late Fusion	183	164	64	30
Color Attention	10	13	36	87

Table 5 Images from bird, pottedplant, motorbike and sofa categories from the PASCAL VOC 2007 data set. The number indicates the rank for the corresponding object category. A lower number reflects higher confidence on the category label. The object category list contains 4952 elements in total. Color attention outperforms SIFT, early and late fusion on the bird, pottedplant and sofa category images. On motorbike category late fusion provides better ranking than color attention.

Without the novel feature selection scheme a mean average precision of 57.5 was reported. A similar experiment was performed by van de Sande et al (2010) where all the color descriptors (C-SIFT, rg-SIFT, OpponentSIFT and RGB-SIFT) were fused with SIFT and spatial pyramid matching to obtain a map of 60.5. Recently, Harzallah et al (2009) obtained a mean average precision of 63.5 by combining object classification and localization scores. A MAP of 64.0 is reported by Zhou et al (2010) using shape alone with a superior coding scheme. This scheme yields a gain of 19.4% over standard vector-quantization used in our framework.

Table 6 shows the results obtained on 2009 PASCAL VOC data set. Our proposed approach outperforms SIFT over all the 20 categories.

For the PASCAL 2009 challenge submission, we further combine the color attention method with additional ColorSIFT (van de Sande et al, 2010), spatial pyramid matching and combining the classification scores with detection results (Harzallah et al, 2009). We follow the classical bag-of-words pipeline where for each image different features are detected. A variety of feature extraction schemes such as GIST (Oliva and Torralba, 2001) are employed afterwards followed by vocabulary and histogram construction. Spatial information is captured using spatial pyramid histograms (Lazebnik et al, 2006) by dividing the image into 2 × 2 (image quarters) and 1 × 3 (horizontal bars) subdivisions. We compressed the visual vocabularies using the agglomerative information bottleneck approach (Fulkerson et al, 2008). Finally, color attention is combined to provide as an input to the classifier. By using SIFT, we obtained a mean average precision (MAP) of 51.0 on the validation set. By adding color attention, we obtained a significant performance gain with a MAP score of 56.2. Finally we added additional descriptors to achieve a MAP of 59.4. Our final submission which also included the object localization results obtained best results on potted plants and tvmonitor category in the competition [10].

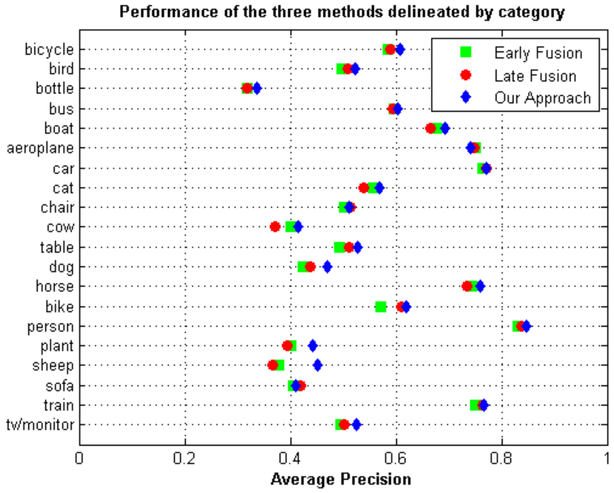

Fig. 8 Results per category on PASCAL VOC 2007 data set: the results are split out per object category. Note that we outperform Early and Late Fusion in 16 out of 20 object categories.

5.7 Caltech-101 Data Set: color and shape co-interference

Finally, our approach is tested in a scenario where combining color with shape has shown to consistently deteriorate the results in literature (Bosch et al, 2007b; Gehler and Nowozin, 2009; Bosch et al, 2007a; Vedaldi

[10] For detailed results on PASCAL VOC 2009, http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2009/results/

Method	Voc Size	Mean AP
$SIFT$	4000	52.1
$TD(SIFT, CN)$	4000, 500	55.1
$TD(SIFT, HUE)$	4000, 300	54.9
$TD(SIFT, \{CN, HUE\})$	4000, {500, 300}	56.1
$CA(SIFT, CN)$	4000, 500	55.6
$CA(SIFT, HUE)$	4000, 300	55.4
$CA(SIFT, \{CN, HUE\})$	4000, {500, 300}	**56.4**

Table 6 Mean Average Precision on PASCAL VOC 2009 dataset. Note that our results significantly improve the performance over the conventional SIFT descriptor.

et al, 2009; Varma and Ray, 2007). Several factors hamper the performance of color features in this data set: low image quality, number of grayscale images (5%), many graphics-based images in different object categories (i.e. garfield, pigeon, panda etc.) and several object categories (i.e. scissors, Buddha etc.) containing the object placed on a variable color background.

The Caltech-101 data set contains 9000 images divided into 102 categories. We followed the standard protocol (Bosch et al, 2007b; Gehler and Nowozin, 2009; Bosch et al, 2007a; Lazebnik et al, 2006) for our experiments by using 30 images per category for the training and upto 50 images per category for testing. Multi-way image classification is obtained by empolying a one-vs-all SVM classifier. A binary classifier is learned to distinguish each class from the rest of the categories. For each test image, the category label of the classifier is assigned that provides the maximum response. We provide results over all 102 categories and the final recognition performance is measured as the mean recognition rate per category.

Method	Voc Size	Score
$SIFT$	500	73.3
$EarlyFusion(SIFT, CN)$	1000	70.6
$LateFusion(SIFT, CN)$	500 + 500	74.9
$OpponentSIFT$	1000	66.3
$C - SIFT$	1000	59.7
$TD(SIFT, CN)$	500, 500	74.7
$CA(SIFT, CN)$	500, 500	**76.2**

Table 7 Recognition results on Caltech-101 Set. Note that conventional early fusion based approaches to combine color and shape provide inferior results compared to the results obtained using shape alone.

Table 7 shows the results obtained using spatial pyramid representations upto level 2. Among the existing approaches, only late fusion provides a gain over shape alone. For all early fusion approaches inferior results are obtained compared to shape alone. Our approach that combines the strength of both early and late fusion improves the recognition performance on this data set. Introducing color information is beneficial for some categories such as flamingo-head, pizza, lobster, dolphin etc. whereas recognition performance of categories such as hedgehog, gramophone, pigeon, emu etc. are hampered by combining color and shape.

In Fig. 9, a performance comparison of early and late fusion versus color attention is given. For all the categories below the diagonal, color attention outperforms early and late fusion. As illustrated in Fig. 9 for most of the object categories in this data set, the best results are obtained using color attention.

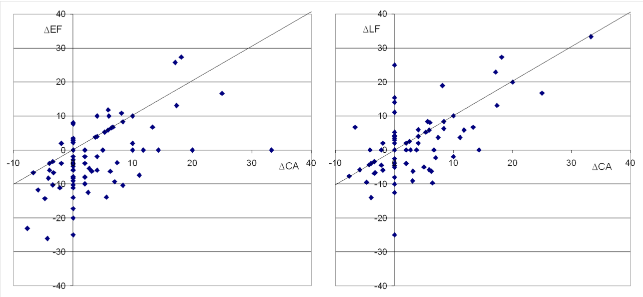

Fig. 9 Left figure: comparison of gain over shape obtained by early fusion (ΔEF) to gain obtained by color attention (ΔCA). Every dot represents one of the Caltech-101 categories. All points above the origin show an advantage of early fusion over shape. All points on the right of origin depict a gain of color attention over shape. For all points below the diagonal color attention outperforms early fusion. Similar results for late fusion are shown in the figure on the right.

The best results reported on this data set is 82.1% by Gehler and Nowozin (2009) using variants of multiple kernel learning to combine 49 different kernel matrices of 8 different types of features such as SIFT, ColorSIFT, HOG, LBP, V1S+ etc. Our proposed approach can be further employed together with previously used features to further boost the results. In Table 8 we compare to other approaches which combine color and shape cues. Note that we do not learn class-specific weights of the spatial pyramid levels which has been shown to improve the results significantly (Bosch et al, 2007b; Gehler and Nowozin, 2009; Bosch et al, 2007a; Vedaldi et al, 2009) mainly due to the fact that objects are always in the center of the image. Results show that early fusion combination of color and shape deteriorates results significantly upto 12%. Our approach improves the overall performance on this data set compared to shape alone.

6 Conclusions

In this paper we have performed an analysis on two existing approaches (early and late fusion) that combine color and shape features. Experimental results clearly

Method	Shape	Color-Shape	Score
Bosch et al (2008)	71.6	68.2	−3.4
Varma and Ray (2007)	52.8	40.8	−12.0
Vedaldi et al (2009)	73.0	63.0	−10.0
Gehler and Nowozin (2009)	66.4	55.0	−11.4
OurApproach	73.3	76.2	**+2.9**

Table 8 Comparison in performance of shape and color-shape approaches reported in literature with our proposed approach. Note that our method improves the overall recognition performance over shape alone on Caltech-101 data set.

demonstrate that both these approaches are sub-optimal for a subset of object categories. This analysis leads us to define two desired properties for feature combination: *feature binding* and *feature compactness*, which in a standard bag-of-words approach are mutually exclusive.

We present a new image representation which combines color and shape within the bag-of-words framework. Our method processes color and shape separately and then combines it by using both bottom-up and top-down attention. The bottom-up component of color attention is obtained by applying a color saliency method whereas the top-down component is obtained by using learned category-specific color information. The bottom-up and top-down attention maps are then used to modulate the weights of local shape features. Consequently, a class-specific image histogram is constructed for each category.

Experiments are conducted on standard object recognition data sets. On the two data sets, Soccer and Flower, where color plays a pivotal role, our method obtains state-of-the-art results increasing classification rate over 5% compared to early and late fusion. On the PASCAL VOC data sets, we show that existing methods based on early fusion underperform for classes with shape-color independency, including many man-made classes. Results based on color attention show that also for these classes color does contribute to overall recognition performance. Performance comparison of our approach to existing fusion approaches has been shown in Table 9.

Data set	SIFT	Early Fusion	Late Fusion	CA
Soccer	50	90	88	96
Flower	69	89	88	95
PASCAL VOC	53.7	55.7	56.0	58.0
Caltech-101	73.3	70.6	74.9	76.2

Table 9 Comparison of our approach with existing fusion approaches on various data sets. Note that our approach outperforms early and late fusion on all data sets.

The dimensionality of color attention histogram is equivalent to the number of object categories times the size of the shape vocabulary. Therefore as a future research direction, we aim to look at dimensionality reduction techniques such as PCA and PLS to reduce the dimensionality of color attention histograms. Another interesting future research line includes looking into other visual features that can be used as an attention cue. Recently, Li et al (2010a,b) have applied our model to incorporate motion features as an attention cue and demonstrated its effectiveness for event recognition. We believe that top-down guidance can also improve the performance in several other applications such as object detection and action recognition.

Acknowledgements

This work is funded by the Project MEC TIN2007-64577, TIN2010-21771-C02-1, MEC TIN2009-14173, 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

Bach F (2008) Exploring large feature spaces with hierarchical multiple kernel learning. In: NIPS

Bosch A, Zisserman A, Munoz X (2006) Scene classification via plsa. In: ECCV

Bosch A, Zisserman A, Munoz X (2007a) Image classification using random forests and ferns. In: ICCV

Bosch A, Zisserman A, Munoz X (2007b) Representing shape with a spatial pyramid kernel. In: CIVR

Bosch A, Zisserman A, Munoz X (2008) Scene classification using a hybrid generative/discriminative approach. PAMI 30(4):712–727

Burghouts GJ, Geusebroek JM (2009) Performance evaluation of local colour invariants. CVIU 113:48–62

Cai H, Yan F, Mikolajczyk K (2010) Learning weights for codebook in image classification and retrieval. In: CVPR

Dorko G, Schmid C (2003) Selection of scale-invariant parts for object class recognition. In: ICCV

Everingham M, Gool LV, Williams CKI, JWinn, Zisserman A (2007) The pascal visual object classes challenge 2007 results.

Everingham M, Gool LV, Williams CKI, Winn J, Zisserman A (2008) The pascal visual object classes challenge 2008 (voc2008) results. [online]. available: http://www.pascal-network.org/challenges/voc/voc2008/

Everingham M, Gool LV, Williams CKI, JWinn, Zisserman A (2009) The pascal visual object classes challenge 2009 results.

Fei-Fei L, Perona P (2005) A bayesian hierarchical model for learning natural scene categories. In: CVPR

Fulkerson B, Vedaldi A, Soatto S (2008) Localizing objects with smart dictionaries. In: ECCV

Gao D, Han S, Vasconcelos N (2009) Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. PAMI 31(6):989–1005

Gehler PV, Nowozin S (2009) On feature combination for multiclass object classification. In Proc. ICCV

Harzallah H, Jurie F, Schmid C (2009) Combining efficient object localization and image classification. In: ICCV

Ito S, Kubota S (2010) Object classification using hetrogeneous co-occurrence features. In: ECCV

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

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

Jurie F, Triggs B (2005) Creating efficient codebooks for visual recognition. In: ICCV

Khan FS, van de Weijer J, Vanrell M (2009) Top-down color attention for object recognition. In: ICCV

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

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

Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proc. CVPR

Li L, Hu W, Li B, Yuan C, Zhu P, Li W (2010a) Event recognition based on top-down motion attention. In Proc. ICPR

Li L, Yuan C, Hu W, Li B (2010b) Top-down cues for event recognition. In: ACCV

Liu T, Sun J, Zheng N, Tang X, Shum H (2007) Learning to detect a salient object. In: CVPR

Lowe DG (2004) Distinctive image features from scale-invariant points. IJCV 60(2):91–110

Marszalek M, Schmid C, Harzallah H, van de Weijer J (2007) Learning object representation for visual object class recognition 2007. In: Visual recognition Challenge Workshop in conjuncture with ICCV

Meur OL, Callet PL, Barba D, Thoreau D (2006) A coherent computational approach to model bottom-up visual attention. PAMI 28(5):802–817

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

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

Nilsback ME, Zisserman A (2006) A visual vocabulary for flower classification. In: CVPR

Nilsback ME, Zisserman A (2007) Delving into the whorl of flower segmentation. In: BMVC

Nilsback ME, Zisserman A (2008) Automated flower classification over a large number of classes. In: ICVGIP

Nowak E, Jurie F, Triggs B (2006) Sampling strategies for bag-of-features image classification. In: ECCV

Oliva A, Torralba AB (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV 42(3):145–175

Orabona F, Luo J, Caputo B (2010) Online-batch strongly convex multi kernel learning. In: CVPR

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

Peters RJ, Itti L (2007) Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In: CVPR

Quelhas P, Odobez JM (2006) Natural scene image modeling using color and texture visterms. In: CIVR

Quelhas P, Monay F, Odobez J, Gatica-Perez D, Tuytelaars T, Gool LV (2005) Modelling scenes with local descriptors and latent aspects. In: ICCV

Rakotomamonjy A, Bach F, Canu S, Grandvalet Y (2007) More efficiency in multiple kernel learning. In: ICML

van de Sande K, Gevers T, Snoek C (2008) Evaluation of color descriptors for object and scene recognition. In: CVPR

van de Sande KEA, Gevers T, Snoek CGM (2010) Evaluating color descriptors for object and scene recognition. PAMI 32(9):1582–1596

Sivic J, Zisserman A (2003) Video google: A text retrieval approach to object matching in videos. In: ICCV

Snoek CGM, Worring M, Smeulders AWM (2005) Early versus late fusion in semantic video analysis. In: ACM MM

Stottinger J, Hanbury A, Gevers T, Sebe N (2009) Lonely but attractive: Sparse color salient points for object retrieval and categorization. In: CVPR Workshops

Treisman A (1996) The binding problem. Current Opinion in Neurobiology 6:171–178

Tsotsos J, SM Culhan and WW, Lai Y, Davis N, Nuflo F (1995) Modeling visual-attention via selective tuning. Artif Intell 78:507–545

Tuytelaars T, Schmid C (2007) Vector quantizing feature space with a regular lattice. In: ICCV

Varma M, Babu BR (2009) More generality in efficient multiple kernel learning. In: ICML

Varma M, Ray D (2007) Learning the discriminative power-invariance trade-off. In: ICCV

Vazquez E, Gevers T, Lucassen M, van de Weijer J, Baldrich R (2010) Saliency of color image derivatives: A comparison between computational models and human perception. Journal of the Optical Society of America A (JOSA) 27(3):1–20

Vedaldi A, Gulshan V, Varma M, Zisserman A (2009) Multiple kernels for object detection. In: ICCV

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

Walther D, Koch C (2006) Modeling attention to salient proto-objects. Neural Networks 19:1395–1407

van de Weijer J, Schmid C (2006) Coloring local feature extraction. In: ECCV

van de Weijer J, Schmid C (2007) Applying color names to image description. In: ICIP

van de Weijer J, Gevers T, Bagdanov AD (2006) Boosting color saliency in image feature detection. PAMI 28(1):150–156

van de Weijer J, Schmid C, Verbeek JJ, Larlus D (2009) Learning color names for real-world applications. IEEE Transaction in Image Processing (TIP) 18(7):1512–1524

Wettschereck D, Aha DW, Mohri T (1997) A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artificial Intelligence Review 11:273–314

Winn JM, Criminisi A, Minka TP (2005) Object categorization by learned universal visual dictionary. In: ICCV

Wolfe JM (2000) The Deployment of Visual Attention:Two Surprises. Search and Target Acquisition, edited by NATO-RTO, NATO-RTO.

Wolfe JM, Horowitz T (2004) What attributes guide the deployment of visual attention and how do they do it? Nature Reviews Neuroscience 5:1–7

Xie N, Ling H, Hu W, Zhang X (2010) Use bin-ratio information for category and scene classification. In: CVPR

Yang L, Jin R, Sukthankar R, Jurie F (2008) Unifying discriminative visual codebook generation with classifier training for object category recognition. In: CVPR

Zhou X, Yu K, Zhang T, Huang TS (2010) Image classification using super-vector coding of local image descriptors. In: ECCV