The Impact of Color on Bag-of-Words based Object Recognition

David A. Rojas Vigo, Fahad Shahbaz Khan, Joost van de Weijer and Theo Gevers
Computer Vision Center, Universitat Autonoma de Barcelona, Spain.
david.rojas@cvc.uab.cat

Abstract

In recent years several works have aimed at exploiting color information in order to improve the bag-of-words based image representation. There are two stages in which color information can be applied in the bag-of-words framework. Firstly, feature detection can be improved by choosing highly informative color-based regions. Secondly, feature description, typically focusing on shape, can be improved with a color description of the local patches. Although both approaches have been shown to improve results the combined merits have not yet been analyzed. Therefore, in this paper we investigate the combined contribution of color to both the feature detection and extraction stages. Experiments performed on two challenging data sets, namely Flower and Pascal VOC 2009; clearly demonstrate that incorporating color in both feature detection and extraction significantly improves the overall performance.

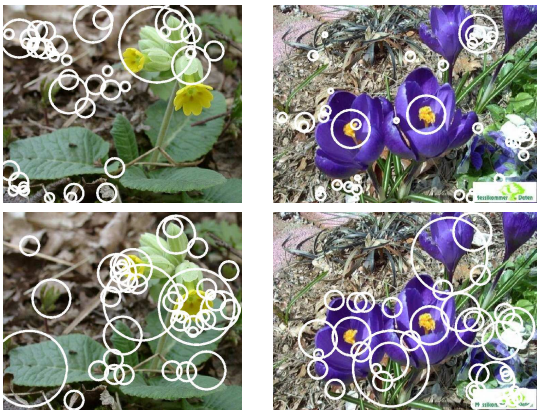

Figure 1: Top row: Laplacian-of-Gaussian and Harris Laplace based on luminance. Bottom row: Laplacian-of-Gaussian and Harris Laplace based on color boosting. The color boosted detector focuses on the more informative features. Only the fifty strongest detected regions are depicted. The radius of the circles indicates the scale selected.

1. Introduction

Put succinctly, the bag-of-words based image representation is currently the most successful approach for object and scene recognition. The first stage, called feature detection, involves region selection. As a next step all selected regions are represented using local descriptors followed by vector quantization into a fixed-size codebook of visual words. Finally each image is represented by a histogram of the visual words. A classifier (typically a non-linear SVM) is then used to recognize the categories based on these histogram representations of the images. In this paper we focus on the first two stages, namely feature detection and extraction, for the incorporation of color information.

The aim of feature detection is to find discriminative regions in images. Most existing methods only use the shape saliency as a criterion for detection, which is extracted from the luminance information. However the use of color information could lead to the detection of more salient regions (see Fig. 1). To date, there have been limited investigations into the usage of color at the detection stage. Recently [8, 1], it has been shown that color can be successfully used to improve the performance of detection. However, none of these studies have evaluated the combined impact of color at both the detection and description levels for object recognition.

Adding color in the description phase of the bag-of-words framework has been more widely studied [11, 9, 6]. Two well-known approaches to fuse color and shape information are early and late fusion. In early fusion, a joint color-shape vocabulary is constructed whereas late fusion concatenates histogram representation of both color and shape, obtained independently.

Recently Fahad et al. [3] proposed a novel way of adding the color information where color is used to guide attention by means of a top-down category-specific attention map employed to modulate the shape words computed using a standard bag-of-words approach. This approach was shown to outperform both

early and late fusion. In such a framework the contribution of color at the feature detection stage is still not investigated. The color attention approach focuses on obtaining a compact image representation by combining the advantages of both early and late fusion schemes. In such a framework the contribution of color at the feature detection stage is still not investigated and ignored.

The review of the existing research shows that color feature detection improves results when using luminance descriptors [8], and that color feature detection combined with color feature description improves results [9]. However, the question of whether or not both are required to obtain optimal results has not yet been answered. Therefore, in this paper, we investigate the performance gain obtained from color in the feature detection stage and in the feature description stage separately. Results show that for optimal results color information should be exploited in both the feature detection and description phase.

2. Color Feature Detection

Although the use of color information is limited by various practical difficulties, the conversion to gray-value has a number of side-effects that are particularly undesirable for local feature detection. It is well known that gray-value versions of color images do not preserve chromatic saliency, i.e. regions that exhibit chromatic variation often lose their distinctiveness when mapped to scalars based on isoluminance. To exploit the color information of images, we need to (1) use a representation that makes the color saliency explicit, and (2) extend multi-scale feature extraction theory from scalars to vectors. In the following we extend three of the most commonly used detectors [5], namely Laplacian-of-Gaussian, Harris-Laplace, and Hessian, to incorporate color saliency.

2.1 Color saliency

The color saliency boosting algorithm has provided an efficient method to exploit the saliency of color edges based on information theory. It was proposed by van de Weijer et al. [10] and it has been successfully applied to image retrieval and image classification [8] [9].

Let $\mathbf{f}$ be a color image and $\mathbf{f}_x = (R_x\ G_x\ B_x)$ its corresponding spatial derivative. The information content of first-order directional derivative in a local neighborhood is given by $I(\mathbf{f}_x) = -log(p(\mathbf{f}_x))$, where $p(\mathbf{f}_x)$ is the probability of the spatial derivative. Note that a derivative has a higher content of information if it has a low probability of occurrence.

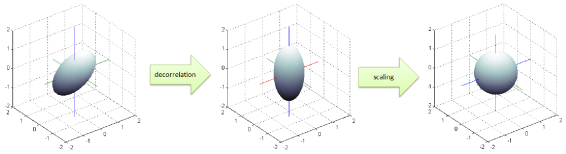

Figure 2: Transformation of color derivatives distribution by the color saliency boosting algorithm. It first decorrelates the original distribution, and then applies a scaling of the axes.

The color derivatives distribution is dominated by a principal axis of maximum variation, caused by the luminance, and two minor axes, attributed to chromatic changes. This means that our image representation assigns a high probability to changes in luminance with respect to color changes. Therefore, the color saliency boosting function $g(.)$ is a linear transformation that makes these probabilities more uniform.

Consider the first-order color derivatives, $\mathbf{f}_x$ which form a zero-mean distribution that can be characterized by its second-order statistics, i.e. its covariance matrix $\Sigma_x = E[\mathbf{f}_x^T \mathbf{f}_x]$, whose eigenvectors define three new axes in which the components of the distribution are decorrelated and can be rescaled. Therefore, this algorithm transforms this distribution into a more homogeneous by applying a decorrelation and whitening:

$$g(\mathbf{f}_x) = \Sigma_x^{-\frac{1}{2}} \mathbf{f}_x, \tag{1}$$

therefore we have a new distribution with the same variance in all directions so that changes in luminance and color have the same impact for feature detection. This is depicted in Fig. 2. Similar equations hold for $\mathbf{f}_y$.

Note that the same influence of luminance changes on the chromatic changes in the image is reflected not only in the derivatives of first order, but also in higher order. Therefore the extension of this theory to higher order operators is straightforward. For example, for second order derivatives we can define the color saliency boosting function as

$$g(\mathbf{f}_{xx}) = \Sigma_{xx}^{-\frac{1}{2}} \mathbf{f}_{xx}, \tag{2}$$

where Σ_{xx} is the covariance matrix of the second-order directional derivatives. Similar equations hold for $\mathbf{f}_{xy}$ and $\mathbf{f}_{yy}$. Another property of this transformation is that due to its linearity it can also be applied to the original image as a preprocessing operation before detection. Fig. 3 illustrates two examples of edge detection.

2.2 From luminance to color

The extension from luminance to color signals is an extension from scalar to vector signals. A basic ap-

Figure 3: Comparison for edge detection. Top row: Original images. Middle row: RGB color edges. Bottom row: color-boosted edges. RGB edges are more biased by luminance.

proach to extend existing detectors to color is to compute the derivatives of each channel separately and then combine the partial results. However, combining the first derivatives with a simple addition of the separate channels results in cancellation in the case of opposing vectors, and the same situation occurs for second-derivative operators. To overcome this problem, Di Zenzo [13] proposed the color tensor defined as

$$\mathbf{G} = \begin{bmatrix} R_x^2+G_x^2+B_x^2 & R_xR_y+G_xG_y+B_xB_y \\ R_xR_y+G_xG_y+B_xB_y & R_y^2+G_y^2+B_y^2 \end{bmatrix} \quad (3)$$

This definition can be considered as a simple extension of the second moment matrix to color, and it has been successfully used to extend first order operators to color. We compute the color Harris-Laplace detector using

$$det(\mathbf{G}) - \alpha trace^2(\mathbf{G}) > threshold \quad (4)$$

where α and $threshold$ are two detector parameters. However, it does not generalize to the second-derivative operators, like the Hessian matrix. Therefore, new methods are required to combine the differential structure of color images in a principled way.

The definition of Laplacian-of-Gaussian operator, among many other state-of-the-art detectors, comes from the Hessian matrix. Thus, in order to extend this operator to color we need a precise mathematical definition of the Hessian matrix for color images, which consider the problem of opposing channels. Shi et al. [7] showed an extension of this matrix to color using a quaternion representation of color images, which overcomes the problem of opposing channels. From this definition it can be demonstrated that it is possible to derive an extension to color by combining channels in a vectorized fashion. Therefore, we extend the Laplacian-of-Gaussian detector to multiple channels combining responses of individual channels using a generalized scale-normalized Laplacian operator defined by

$$Color\ LoG(\sigma) = \sigma^2 \|(\mathbf{f}_{xx} + \mathbf{f}_{yy})\| \quad (5)$$

And we can define the Color Hessian detector by

$$Color\ Hessian(\sigma) = \sigma^2 \|(\mathbf{f}_{xx}\mathbf{f}_{yy} - \mathbf{f}_{xy}{}^2)\| \quad (6)$$

where $\mathbf{f}_{xx}$, $\mathbf{f}_{yy}$ and $\mathbf{f}_{xy}$ are the second-order directional derivatives with scale σ, and $\|\cdot\|$ is the vector norm. This extension leads to a scale-space representation which includes the contributions of luminance and chromatic components in a scalar-valued representation. In Fig. 1 the results of color-boosted Laplacian-of-Gaussian and Harris-Laplace are provided. It is shown that color-boosted detectors capture more informative regions as compared to luminance-only detectors [1].

3. Color Feature Description

To apply color in the feature description step, we use the Color Attention method [3] which was shown to obtain state-of-the-art results.

The human visual system has a remarkable ability of reducing the computational cost of a data-driven visual search by means of an attentional mechanism. The two distinctive ways of directing attention are *bottom-up*, which involves detecting salient regions in an image used for the deployment of visual attention and, *top-down*, which makes use of prior knowledge available for a specific target to guide the visual attention. It is further believed that basic features of visual objects are processed separately before they are combined in the presence of attention for the final representation. Similarly, in our framework, color and shape are processed separately before they are combined by means of attention. This attention is top-down in nature and is guided by the color feature (attention cue) of visual objects. The attention cue describes our prior knowledge about the categories we are looking for and is further deployed to weight the shape features (descriptor cue).

[1] Code of the color-based detectors proposed here is available on-line: http://cat.cvc.uab.es/ daro-jas/FeatureDetection/FeatureDetection.html

Within the bag-of-words framework each image I_i, i=1,2,...,N contains a number of detected local features f_{ij}, j=1,2,...,M^i . These local features are then represented by the visual words w_i^k, i=1,2,...,V^k and $k \in \{s, c\}$ for the two cues, shape and color. The computation of top-down color attention based image representation is done according to:

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

The probabilities $p\left(class|w_{ij}^c\right)$ are computed by using Bayes,

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

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

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

obtained by summing over the indexes to the training images of the category I^{class}. We use the training data for computing the uniform prior over the classes $p\left(class\right)$. By computing $p\left(class|w_{ij}^c\right)$ for all local features in an image, a class-specific color attention map is constructed. This map is used to modulate the sampling of shape features; in regions with high attention more shape features are sampled than in regions with low attention. This leads to a different distribution over the same shape visual words for each category. The final image representation is obtained by concatenating all the class-specific histograms.

4. Experiments

Here we analyze the relative performance gain obtained while using color in the feature detection and description phase for object recognition.

Implementation Details: In our experiments we use the detectors proposed in section 2 together with a multi-scale Grid. The SIFT descriptor is used to create a shape vocabulary and the Color Names and HUE descriptors are used to create color vocabularies. We abbreviate our results with the notation convention CA (descriptor cue, attention cues) where CA stands for Color Attention based bag-of-words. In our experiments we use a standard nonlinear SVM with a χ^2 kernel for the Flower data set and intersection kernel [4] for the Pascal VOC 2009 data set since it requires significantly less computational time, while providing performance similar to that using a χ^2 kernel. We tested our approach on two different and challenging data sets,

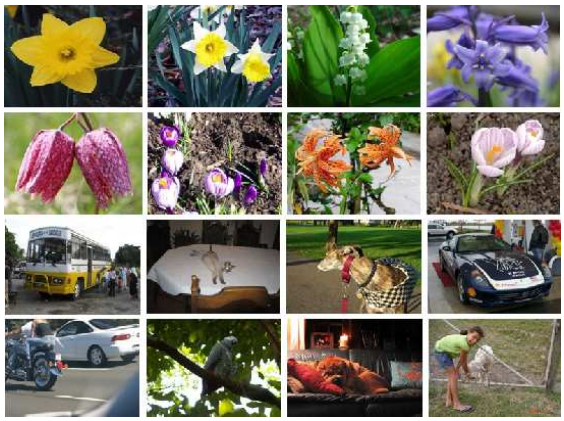

Figure 4: Example from the two data sets The top two rows show images from the Flower data set and the bottom two rows provide examples from the Pascal VOC 2009 data set.

namely Flower and PASCAL VOC Challenge 2009. The Flower data set [2] consists of 17 classes of different varieties of flower species and each class has 80 images, divided into 60 training and 20 test images. The PASCAL VOC 2009 data set [3] is currently the benchmark in image classification. It consists of 13704 images of 20 classes with 7054 training images and 6650 test images. Fig. 4 shows some of the images from these two data sets.

Oxford Flower Set: Image classification results on the Flower data set illustrate the performance of our approach on a data set where both color and shape features are important as some flowers are clearly distinguished by shape, e.g. daisies and some by color, e.g. fritillaries.

Method	Detector	Score
$SIFT$	$Intensity$	72
$SIFT$	$Boosted$	73
$SIFT$	$Both$	73
$CA(SIFT, \{CN, HUE\})$	$Intensity$	85
$CA(SIFT, \{CN, HUE\})$	$Boosted$	89
$CA(SIFT, \{CN, HUE\})$	$Both$	**89**

Table 1: Classification Score on Flower Data set.

The results in Table 1 clearly demonstrate the contribution of color in the feature detection phase. The color boosted interest point sampling strategies significantly outperform the color attention results obtained by using only intensity information in feature detection. Combining intensity-based and color-boosted feature detection provides the same results as using only color-boosted feature detectors. Our approach leads to state-of-the-art results on this data set comparable with

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

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

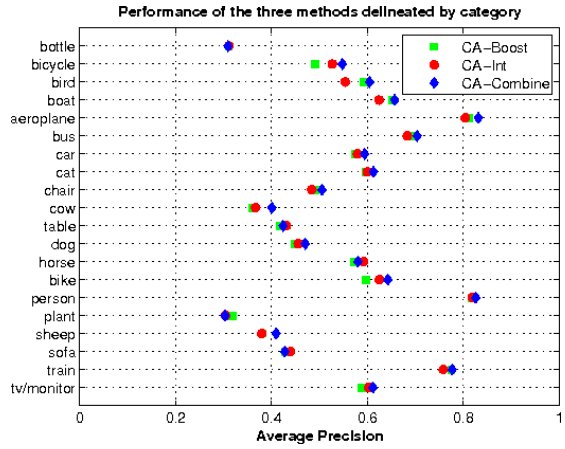

Figure 5: Results per category on Pascal 2009 data set: the results are split out per object category. Note that the combination of intensity and color-boosted detection along with color attention provides the best results. But on categories such as potted-plants color-boosted detectors provide the best results.

the ones previously reported in literature [6, 2, 3, 12].

Pascal VOC 2009: Finally, we present results on a data set where the shape cue is predominant and color plays a subordinate role. For this data set the average precision is used as a performance metric in order to determine the accuracy of recognition results.

Method	Detector	Mean AP
$SIFT$	$Intensity$	51.3
$SIFT$	$Boosted$	49.8
$SIFT$	$Both$	52.1
$CA(SIFT, \{CN, HUE\})$	$Intensity$	54.7
$CA(SIFT, \{CN, HUE\})$	$Boosted$	54.6
$CA(SIFT, \{CN, HUE\})$	$Both$	**56.1**

Table 2: Pascal VOC 2009 MAP Scores.

Table 2 shows the results for both SIFT and Color Attention. The results (SIFT and CA) based on intensity-based point detection are slightly better than those of color-boosted detection schemes. This is due to the fact that this data set is predominantly shape-based and color alone is not a substantial cue. However the combination of intensity and color-boosted detection significatively improves the results suggesting that the convincing gain is owing to the complementary nature of both these detection schemes. The results per category are presented in Fig. 5.

Our final submission for the 2009 Pascal competition involves an extension of the framework proposed in this paper with more descriptors. Further, combining the framework with object localization scores led to results that are very close to state-of-the-art.

5. Conclusion

We presented an analysis on how to optimally apply color in the bag-of-words approach to image classification. The outcome of our experiments show that color should be used both in the feature detection and the feature extraction stages. In particular, we show that color feature detection does further improve image classification results based on the color attention approach.

6. Acknowledgments

This work is partially supported by the Ramon y Cajal program, Consolider-Ingenio 2010 CSD2007-00018 and TIN2009-14173 of Spanish Ministry of Science, and the Marie Curie Reintegration program, TS-VICI224737 of the European Union.

References

[1] P.-E. Forssén. Maximally stable colour regions for recognition and matching. In *CVPR*, 2007.

[2] P. V. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, 2009.

[3] F. S. Khan, J. van de Weijer, and M. Vanrell. Top-down color attention for object recognition. In *ICCV*, 2009.

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

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

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

[7] L. Shi, B. Funt, and G. Hamarneh. Quaternion color curvature. In *IS&T Sixteen Color Conference*, 2008.

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

[9] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 2010.

[10] J. van de Weijer, T. Gevers, and A. Bagdanov. Boosting color saliency in image feature detection. *PAMI*, 28:150–156, 2006.

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

[12] N. Xie, H. Ling, W. Hu, and Z. Zhang. Use bin-ratio information for category and scene classification. In *CVPR*, 2010.

[13] S. D. Zenzo. A note on the gradient of a multi-image. *Computer Vision, Graphics and Image Processing*, 33:116–125, 1986.