# Color Descriptors for Object Category Recognition

***Koen E.A. van de Sande, Theo Gevers and Cees G.M. Snoek; University of Amsterdam; Amsterdam, The Netherlands***

## Abstract

*Category recognition is important to access visual information on the level of objects. A common approach is to compute image descriptors first and then to apply machine learning to achieve category recognition from annotated examples. As a consequence, the choice of image descriptors is of great influence on the recognition accuracy. So far, intensity-based (e.g. SIFT) descriptors computed at salient points have been used. However, color has been largely ignored. The question is, can color information improve accuracy of category recognition?*

*Therefore, in this paper, we will extend both salient point detection and region description with color information. The extension of color descriptors is integrated into the framework of category recognition enabling to select both intensity and color variants. Our experiments on an image benchmark show that category recognition benefits from the use of color. Moreover, the combination of intensity and color descriptors yields a 30% improvement over intensity features alone.*

## Introduction

Automatic object category recognition is important to access visual information on the level of objects (buildings, cars, *etc*.). A common approach for systems in image retrieval [6, 13, 15, 16] is to compute image descriptors first and then to apply machine learning to achieve object recognition from annotated examples. As a consequence, the choice of image descriptors is of great influence on the recognition accuracy. So far, intensity-based (e.g. SIFT) descriptors [8, 16] computed at salient points have been used. However, color plays an important role in distinguishing different types of objects. Therefore, we believe the addition of color information can boost performance.

The aim of this paper is to study the influence of color by extending both salient point detection and region description with color. First, the Harris corner detector [5] and Laplacian scale selection are extended for multiple channels. Then, the saliency of image features is increased by applying color saliency boosting [14]. For region description, we consider several color extensions of the state-of-the-art SIFT descriptor [8], which is based on intensity information only. Both salient point detection and region description are extended with color information according to the following criteria: 1. invariance to illumination 2. high discriminate power 3. easy to compute. We propose to extend the object recognition framework with color descriptors, which enables the framework to select both intensity and color variants. We hypothesize that color information improves the accuracy of category recognition.

To evaluate the performance of our color extensions, we use a widely adopted benchmark for object category recognition, the PASCAL VOC Challenge 2007 [2]. This benchmark consists of nearly 10,000 photographs, containing one or more of the 20 object categories defined on this dataset. See figure 1 for an overview of the evaluated object categories.

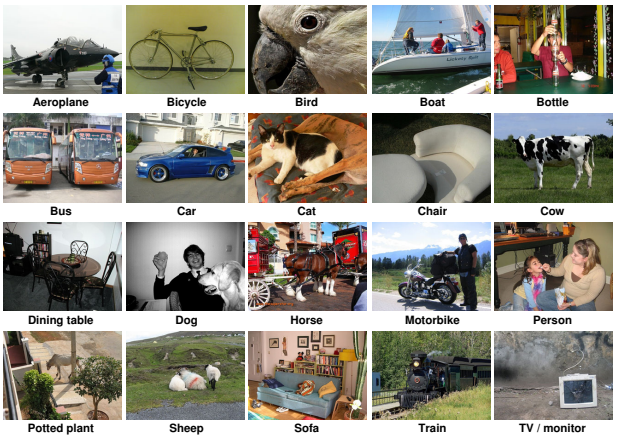


***Figure 1.*** *Object categories of the PASCAL VOC Challenge 2007.*

## Point Detectors

In this section, two scale invariant salient point detectors are discussed: Harris-Laplace and ColorHarris-Laplace with color boosting. Both are based on the Harris corner detector and use Laplacian scale selection [11].

### *Harris Corner Detector*

To extend the Harris corner detector to color information, the intensity-based version is taken. The adapted second moment matrix for position $x$ is defined as follows [10]:

$$\mu(x,\sigma_I,\sigma_D) = \sigma_D^2 g(\sigma_I) \begin{pmatrix} L_x^2(x,\sigma_D) & L_xL_y(x,\sigma_D) \\ L_xL_y(x,\sigma_D) & L_y^2(x,\sigma_D) \end{pmatrix}, \quad (1)$$

with $\sigma_I$ the integration scale, $\sigma_D$ is the differentiation scale and $L_z(x,\sigma_D)$ the derivative computed in the $z$ direction at point $x$ using differentiation scale $\sigma_D$. The matrix describes the gradient distribution in the local neighborhood of point $x$. Local derivatives are computed by Gaussian kernels with a differentiation scale denoted by $\sigma_D$. The derivatives are averaged in the neighborhood of point $x$ by smoothing with a Gaussian window suitable for the integration scale $\sigma_I$.

The eigenvalues of the second moment matrix represent the two principal signal changes in the neighborhood of a point. *Salient points* are extracted where both eigenvalues are significant i.e. the signal change is significant in orthogonal directions, which is true for corners, junctions, etc.

The Harris corner detector [5] relies on the properties of the second moment matrix. It combines the trace and the determinant of the matrix into a *cornerness measure*:

$$cornerness = \det(\mu(x,\sigma_I,\sigma_D)) - \kappa \mathrm{trace}^2(\mu(x,\sigma_I,\sigma_D)), \quad (2)$$

with $\kappa$ an empirical constant with values between 0.04 and 0.06. Local maxima of the cornerness measure (equation 2) determine the salient point locations.

### *Scale Selection*

Automatic scale selection allows for the selection of the characteristic scale of a point, which depends on the local struc-

ture around the point. The characteristic scale is the scale for which a given function attains a maximum over scales. It has been shown [9] that the cornerness measure of the Harris corner detector rarely attains a maximum over scales. Thus, it is not suitable for selecting a proper scale. However, the Laplacian-of-Gauss (LoG) does attain a maximum over scales. Therefore, it will be used in this paper. With $\sigma_n$, the scale parameter of the LoG, it is defined for a point $x$ as:

$$|LoG(x,\sigma_n)| = \sigma_n^2 |L_{xx}(x,\sigma_n) + L_{yy}(x,\sigma_n)|. \quad (3)$$

The function reaches a maximum when the size of the kernel matches the size of the local structure around the point.

### Harris-Laplace Detector

The Harris-Laplace detector uses the Harris corner detector to find potential scale-invariant salient points. It selects a subset of these points for which the LoG reaches a maximum over scale. Mikolajczyk and Schmid [10] define an iterative version of the Harris-Laplace detector and a simplified version which does not involve iteration. The simplified version performs a more thorough search through the scale space by using smaller intervals between scales. The iterative version relies on its convergence property to obtain characteristic scales. Both versions give comparable results on our dataset, so we have chosen the simplified version for the Harris-Laplace detector.

### Color Boosting

From information theory it is known that rare color transitions in an image are very distinctive. By adapting the Harris detector to the saliency of an image, the focus of the detector shifts to more distinctive points. The transformation of the image to achieve this is called color saliency boosting. We use the color boosting transformation in the opponent color space [14]. The opponent space is given by:

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix}. \quad (4)$$

The third channel $O_3$ is equal to the intensity channel of the HSV color model, subject to a scaling factor. $O_1$ and $O_2$ contain the red-green and yellow-blue opponent pairs.

The color boosting transformation is a weighing of the individual opponent channels: $(0.850 \cdot O_1, 0.524 \cdot O_2, 0.065 \cdot O_3)^T$, where the sum of the squared weights is equal to 1. Note that these weights are focused on the red-green and yellow-blue opponent pairs, with almost no weight given to the intensity channel $O_3$.

### ColorHarris-Laplace Detector

To extend the Harris detector to multiple channels $m$, a vector-based version is obtained where $L_x$ is substituted by a vector $\vec{f}_x = \left(L_x^{C_1}, L_x^{C_2}, \ldots, L_x^{C_m}\right)^T$ with $C_i$ the $i^{th}$ channel. The product between derivatives is substituted by a vector inproduct. If the vector is 1-D (*e.g.* an intensity image), this is equivalent to the original second moment matrix. The second moment matrix for the ColorHarris corner detector is:

$$\mu_{ColorHarris}(x,\sigma_I,\sigma_D) = \sigma_D^2 g(\sigma_I) \begin{pmatrix} \vec{f}_x(x,\sigma_D)\cdot\vec{f}_x(x,\sigma_D) & \vec{f}_x(x,\sigma_D)\cdot\vec{f}_y(x,\sigma_D) \\ \vec{f}_x(x,\sigma_D)\cdot\vec{f}_y(x,\sigma_D) & \vec{f}_y(x,\sigma_D)\cdot\vec{f}_y(x,\sigma_D) \end{pmatrix}, \quad (5)$$

with $\vec{f}_x = \left(L_x^{C_1}, L_x^{C_2}, \ldots, L_x^{C_m}\right)^T$ for an image with channels $\{C_1, C_2 \ldots C_m\}$, with $L_x^{C_i}$ being the Gaussian derivative of the $i^{th}$ image channel $C_i$ in direction $x$. The image channels $C_i$ can be instantiated to channels of any color model, such as *RGB*, opponent color model, etc. It is also possible to first apply preprocessing operations (such as color boosting) to an image and instantiate the channels afterwards. For our ColorHarris corner detector, we instantiate the channels to the *R*, *G* and *B* channels of the standard *RGB* color space. We preprocess images using color boosting as discussed above.

The LoG kernel is extended to operate on multiple channels by summing the $m$ individual channels:

$$|LoG(x,\sigma_n)| = \sum_{i=1}^{m} |LoG_{C_i}(x,\sigma_n)|. \quad (6)$$

## Region Descriptors

In this section we will discuss the region descriptors used to describe the image region around the points obtained using the point detectors from the previous section. Our main goal in this section is to extend region description with color information, as to improve the discriminate power of region descriptors.

### SIFT

SIFT as originally proposed by Lowe [7] consists of both a scale invariant point detector and a region descriptor. The detector gives results similar to the Harris-Laplace detector: scale invariant points are detected on corners in an image. We will use the SIFT descriptor only, which describes the local shape of the region around the salient point using edge histograms. It uses information from the intensity channel only. We compute the SIFT descriptor with a 4x4 grid and 8 bins.

### OpponentSIFT

The SIFT descriptor uses only information from the intensity channel. A natural extension is to include the opponent color space. In this way, we decompose the opponent color space into three channels (equation 4), each described using a SIFT descriptor. The information in the $O_3$ channel is equal to the intensity information, while the other channels describe the color information in the image. However, these other channels do contain some intensity information: hence they are not invariant to changes in light intensity. We term this descriptor OpponentSIFT.

### WSIFT

In the opponent color space the red-green and yellow-blue channels ($O_1$ and $O_2$) still contain some intensity information. To add invariance to intensity changes, [3] proposes the W invariant which eliminates the intensity information from these channels. Therefore, the description of these two channels is invariant to light intensity changes. This descriptor is termed WSIFT [1].

### rgSIFT

In the normalized RGB color model, the chromacity components $r$ and $g$ describe the color information in the image, while being invariant to light intensity changes, shadows and shading [4]. For the *rg*SIFT descriptor, we add descriptors for the $r$ and $g$ chromaticity components.

## Experimental Setup

The PASCAL Visual Object Classes (VOC) Challenge [2] provides a yearly benchmark for comparison of object classification systems. Evaluation outside the yearly cycle is possi-

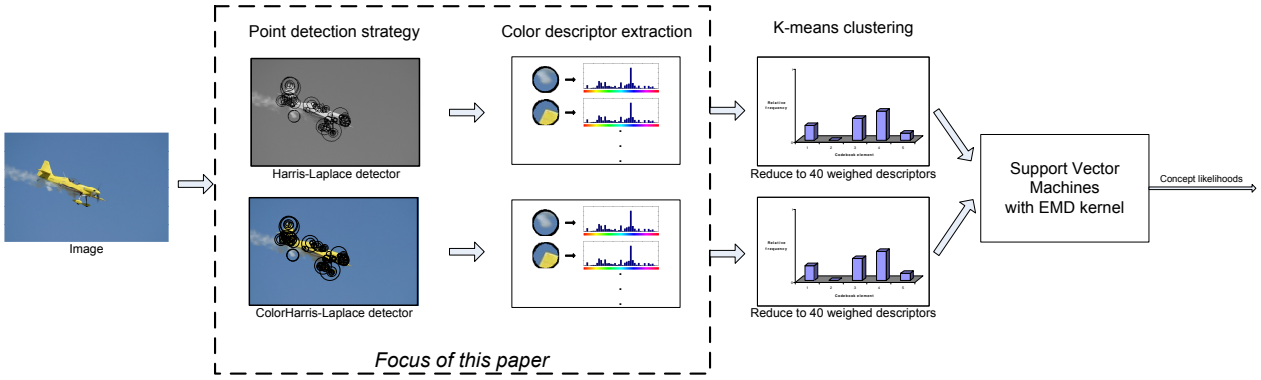


***Figure 2.*** *Overview of the object category recognition pipeline. In the first stage, points are detected in the image, using either Harris-Laplace or ColorHarris-Laplace. In the color descriptor extraction stage, color features are extracted around every sampled point. Next, the color descriptors of an image are reduced to a set of 40 descriptors using $k$-means clustering. These 40 descriptors are retained, together with a weight indicating the number of original descriptors closest to them. The cluster sets for every image form the input to the SVM classifier, which outputs an object category likelihood score for the image. Combinations of different color features are possible during the learning stage. The focus of this paper lies on the point detection strategy and the color descriptors.*

ble because the whole dataset and ground truth annotations are made publicly available. The PASCAL VOC Challenge 2007 dataset contains 10,000 images of 20 different object categories, e.g. bird, bottle, car, dining table, motorbike and people. See figure 1 for a complete overview of the object categories.

Our experimental pipeline, shown in figure 2, learns object appearance models from region descriptors. In the first stage, points are detected in the image, using either Harris-Laplace or ColorHarris-Laplace. In the color descriptor extraction stage, color features are extracted around every sampled point. Next, the color descriptors of an image are reduced to a set of 40 descriptors using $k$-means clustering, to significantly speedup processing in the rest of the pipeline. These 40 descriptors are retained, together with a weight indicating the number of original descriptors closest to them. The Earth Movers Distance (EMD) [12] has been shown to be very suitable for measuring the similarity between cluster sets [16]. The EMD distance between the cluster sets of different images is used in Support Vector Machine (SVM) learning algorithm. To incorporate the EMD distance $D(S_1,S_2)$ between two cluster sets $S_1$ and $S_2$ into the SVM, the distance needs to be transformed into the EMD kernel:

$$K(S_1,S_2) = exp\left(-\frac{1}{A}D(S_1,S_2)\right), \tag{7}$$

where $A$ is a normalization factor equal to the mean value of the EMD distances of all images. Combinations of features can be constructed by summing their normalized EMD distances. To constrain this sum to lie between 0 and 1, it should be divided by the number of features combined. The trained SVM classifier outputs object category likelihood scores.

To allow for a fair comparison between point detectors, we assure that both point detectors output, on average, the same number of points per image.

## Results

We show the overall performance of the intensity point detector, Harris-Laplace, its color extension, ColorHarris-Laplace, and the intensity-based SIFT descriptor and several color descriptors in figure 3. We observe that ColorHarris-Laplace does not perform better than Harris-Laplace. However, given that ColorHarris-Laplace triggers on very different structures in the image, it should be complementary to Harris-Laplace in terms

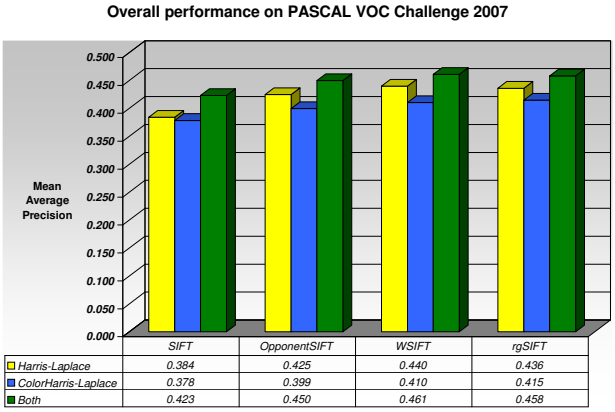


***Figure 3.*** *Mean Average Precision performance over all 20 object categories on the PASCAL VOC Challenge 2007 dataset.*

of discriminative power. Our results using combinations of both Harris-Laplace and ColorHarris-Laplace, also shown in figure 3, confirm this. The combination of point detectors outperforms the individual detectors.

From the results for the different color descriptors, we observe that the color SIFTs perform better than normal intensity-based SIFT. However, the question is how this maps onto the individual object categories: which objects need color? Therefore, we depict in figure 4 a detailed view of figure 3 over individual object categories. Only the best results from figure 3 are shown, *i.e.* the results for the combination of point detectors.

From figure 4, we derive that color descriptors provide a clear improvement for bird, dining table, horse, motorbike, person and potted plant. We observe that the green surroundings of the object (tree leaves or grass) discriminate between false positives and real objects, especially for birds and horses. For dining tables, motorbikes, people and potted plants the color of the object itself is discriminative: furniture is brown, red motorbikes are common, people have similar skin color and potted plants are mostly green. From these results, we see that the best descriptor depends on the object category.

Given that the best descriptor depends on the object category, we believe a descriptor selection strategy on a per-category basis could improve performance further. Our belief is strength-

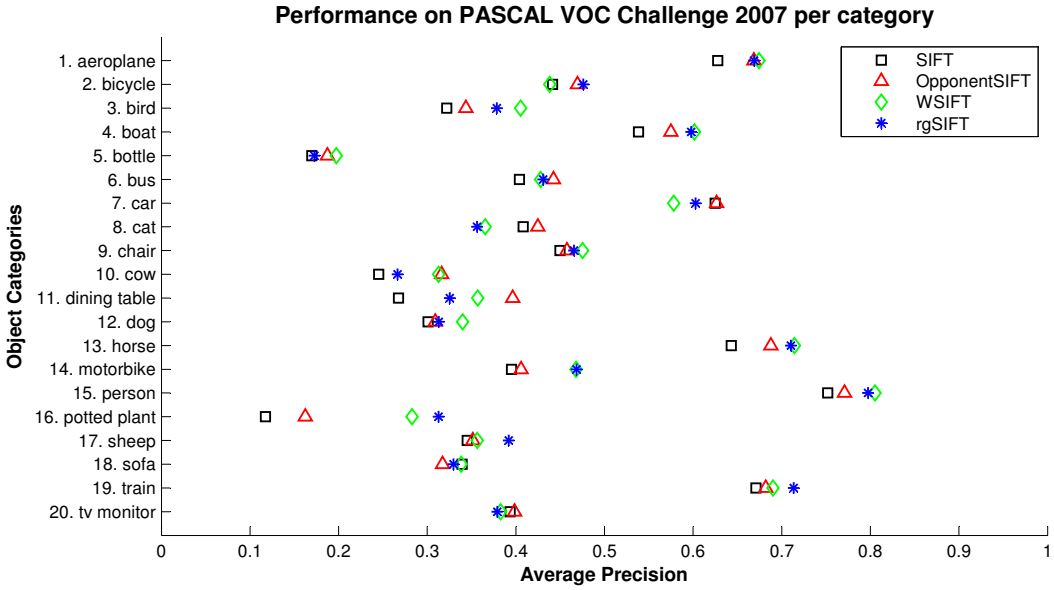


***Figure 4.*** *Average Precision performance of the object categories on the PASCAL VOC 2007 dataset for different color descriptors using the combination of point detectors, consisting of both Harris-Laplace and ColorHarris-Laplace. This combination of point detectors was the best in figure 3.*

end by an experiment using our simple combination scheme. In this experiment, a combination of all color descriptors using both Harris-Laplace and ColorHarris-Laplace is used, resulting in a MAP of 0.503. This is an improvement of 30% over Harris-Laplace with intensity-based SIFT alone.

## Conclusion

In this paper we study the influence of color on salient point detection and description for the purpose of object recognition in images. Our experiments on a real-world image dataset show that object recognition benefits from the use of color. Using color in point detection and color region descriptors yields a 30% performance improvement over using intensity information only.

## Acknowledgments

This work was sponsored by the European Commission through the VIDI-Video project, a sixth framework project (FP6).

## References

[1] G. J. Burghouts and J. M. Geusebroek. Performance evaluation of local color invariants. *Computer Vision and Image Understanding*, 2008. Accepted for publication.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. `http://www.pascal-network.org/challenges/VOC/voc2007/`.

[3] J. M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts. Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338–1350, 2001.

[4] T. Gevers, J. van de Weijer, and H. Stokman. *Color image processing: methods and applications: color feature detection: an overview*, chapter 9, pages 203–226. CRC press, 2006.

[5] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of The Fourth Alvey Vision Conference, Manchester*, pages 147–151, 1988.

[6] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, New York, USA, 2006.

[7] D. G. Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision*, pages 1150–1157, 1999.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[9] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *IEEE International Conference on Computer Vision*, pages 525–531, 2001.

[10] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[11] 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.

[12] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *IEEE International Conference on Computer Vision*, pages 59–66, 1998.

[13] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluation of color descriptors for object and scene recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, 2008.

[14] J. van de Weijer and T. Gevers. Boosting saliency in color image features. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 365–372, San Diego, USA, 2005.

[15] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, C. G. M. Snoek, and A. W. M. Smeulders. Robust scene categorization by learning image statistics in context. In *CVPR Workshop on Semantic Learning Applications in Multimedia (SLAM)*, 2006.

[16] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

## Author Biography

*Koen E.A. van de Sande received a BSc in Computer Science (2004), a BSc in Artificial Intelligence (2004) and a MSc in Computer Science (2007) from the University of Amsterdam. Since then he has worked on his PhD at the University of Amsterdam.*