Improving FREAK Descriptor for Image Classification

Cristina Hilario Gomez, N. V. Kartheek Medathati, Pierre Kornprobst, Vittorio Murino, Diego Sona

Improving FREAK Descriptor for Image Classification

Cristina Hilario Gomez[1](✉), Kartheek Medathati[2], Pierre Kornprobst[2], Vittorio Murino[1], and Diego Sona[1]

[1] Pattern Analysis and Computer Vision, (PAVIS), Istituto Italiano di Tecnologia, Genova, Italy,
{cristina.hilario,vittorio.murino,diego.sona}@iit.it
[2] Neuromathcomp Project Team, INRIA, Sophia Antipolis, France
{kartheek.medathati,pierre.kornprobst}@inria.fr

Abstract. In this paper we propose a new set of bio-inspired descriptors for image classification based on low-level processing performed by the retina. Taking as a starting point a descriptor called FREAK (Fast Retina Keypoint), we further extend it mimicking the center-surround organization of ganglion receptive fields.To test our approach we compared the performance of the original FREAK and our proposal on the 15 scene categories database. The results show that our approach outperforms the original FREAK for the scene classification task.

Keywords: bio-inspired descriptor, binary descriptor, center-surround ganglion cell organization, FREAK, scene classification

1 Introduction

Image classification is a challenging task in computer vision which can be accomplished with a number of different approaches. In particular, scene categorization strongly relies on the appropiate image representation. In the literature, the vast majority of the works use descriptors based on visual information and the recognition of the scenes is achieved either based on the global information or the objects in the image [4][6][7][8]. For example, SIFT [4] and GIST [8] are two of the most used descriptors for scene categorization. Although SIFT was originally proposed for object recognition, it can be used to describe the global features in an image using a bag-of-words (BOW) approach. Under this approach, descriptors are quantized to form a visual codebook. In [4], the authors incorporated spatial information to further improve the BOW model based on SIFT descriptors.

On the other hand, GIST was proposed in [8] as a descriptor based on how humans recognize a scene. Using global information about the scene can significantly improve the classification results. This descriptor, is based on the spatial envelope that represents the most relevant global structure of a scene [7]. Recently, it has been shown that better performance is achieved when both local and global structures in an image are considered [9][17]. In this regard, Census Transform Histogram (CENTRIST) descriptor has been proposed [17], which is based on local binary patterns (LBP) and captures both kind of information.

An alternative approach is to perform image classification inspired by the human visual system. FREAK (Fast Retina Keypoint) was proposed as a fast keypoint descriptor inspired by the retina [1]. The organization of the retina is imitated, using a circular grid where receptive fields of different sizes are considered. The difference in intensity between pairs of receptive fields is calculated and further codified in a binary vector. In particular, the concentration of receptive fields is higher near the center of the pattern, corresponding to the fovea in the retina. In addition to this, they overlap sampling regions adding redundancy which is also present in the retina and this increases the final descriptor discriminative power. FREAK has been evaluated on a matching task showing high object detection performance. BRISK [5] and DAISY [10] are previous descriptors that also compare pairs of intensities using a circular pattern. Compared to state of the art descriptors, such as SIFT, SURF or BRISK, it outperforms them while being faster and simpler. In a new descriptor called CS-FREAK [14], the original grid is simplified reducing the number of receptive fields, and the neighborhood intensity is encoded improving the matching accuracy. In a different kind of task, FREAK has been applied to action recognition in videos through an extension to the descriptor that encodes motion named as MoFREAK [15].

However, biologically inspired descriptors have mainly been applied to object recognition task [1] [5] [10]. In [13] a Difference of Gaussian (DoG) filtering which simulates the performance of the retina is applied to texture classification. In this work, we propose a new set of bio-inspired descriptors for the scene categorization task. Using FREAK descriptor as a baseline, we further enrich it imitating models of the retina. Our proposal is to use a grid based on the center-surround organization of the ganglion receptive fields and perform low-level features extraction in order to classify scenes. In particular, we propose to imitate the ON and OFF cell response by calculating Difference of Gaussians (DoG) of different sizes. Moreover, each receptive field in our grid is described with a linear-nonlinear model (LN) which is typically used in retina models.

The rest of the paper is organized as follows. Section 2 explains the retinal sampling pattern configuration used and describes the construction of each of the descriptors. It also introduces the BOW pipeline used for the classification of the scenes. In Section 3 experimental results on the 15 scene categories dataset are reported. Finally, in Section 4, conclusions are drawn.

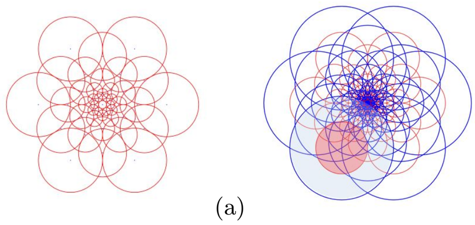
2 Method

In this section we introduce a new set of image descriptors based on the center-surround organization of the ganglion receptive fields.We propose three different binary descriptors each constructed considering different components of ganglion cell response. To start with, the main aspects of FREAK descriptor [1] which are related to our contribution are presented. Next, each of our proposed descriptors are explained in detail. Finally, the bag-of-words approach used for the scene categorization task is introduced.

2.1 Retinal ganglion cells configuration

In FREAK [1], the sampling grid shown in figure 1 is proposed. Each circle corresponds to a receptive field of a ganglion cell and its size represents the standard deviation of the Gaussian kernels applied to the underlying sampling point. However, the Gaussian smoothing applied to each receptive field is approximated calculating the mean intensity. They experimentally observed that changing the size of receptive fields with respect to the log-polar pattern improved the performance. In addition to this, overlapping the receptive fields further improved the results. Based on such a sampling grid, they compared the mean intensity of pairs of receptive fields.

In our model, the configuration of the receptive fields is inspired by FREAK but including several changes to constrain it more closely to biology. As in FREAK, we also consider 43 cells organized in 8 different concentric circles, as can be seen in figure 1.



(a) (b)

Fig. 1. A. Original FREAK sampling pattern. B. Our sampling pattern. In red the center of each receptive field is depicted and in blue the surround. The size of the center corresponds to the size of the cells in the original pattern.

As opposed to FREAK where, for each receptive field (RF) mean averages are calculated, we propose to use a difference of Gaussians (DoG) centered in each cell. Mimicking biology, each RF in our model is composed of a center and a surround. The DoG can be calculated as the subtraction of two Gaussians fitted to each area. More specifically, the radius of the center r_C will be considered as the size of each cell in the original FREAK. The standard deviation of the corresponding Gaussian can be approximated as follows:

$$\sigma_C = \frac{r_C}{3} \, . \tag{1}$$

Based on the literature [16], the relative surround extent has a wide range across retinal ganglion cells. We have empirically chosen the surround to be double size of the center, since the DoG behaves as an edge detector and this is the functionality we are interested in. Therefore, the standard deviation of the surround can be obtained as follows:

$$\sigma_S/\sigma_C = 2 \ . \tag{2}$$

From this equation, the size of the surround can be obtained replaicing equation 2. The formula for the difference of Gaussians (DoG) is the following:

$$K(x,y) = w_C G_{\sigma_C}(x,y) - w_S G_{\sigma_S}(x,y) \ . \tag{3}$$

where w_C and w_S are constants, which determine the type of features estimated by this filtering stage. As before, this parameter presents a high variability depending on the cell type. In our case, the relative surround weight used is $w_S/w_C = 0.9$ based on previous work [16].

2.2 Retinal inspired descriptors

Our descriptors will be estimated based on the retinal activity defined by a classical linear-nonlinear (LN) model, where the activity A of a ganglion cell centered at position (x,y) is defined by:

$$A^{\varepsilon} = N(RF^{\varepsilon}) \quad \text{where for each cell} \qquad RF^{\varepsilon} = \varepsilon I * K(x,y) \ . \tag{4}$$

I is the still image or stimulus and K is the weighted difference of Gaussians. ON and OFF ganglion cells are simulated by setting the parameter ε to respectively $+1$ or -1. The static nonlinear function N is defined by:

$$N(RF^{\varepsilon}) = \begin{cases} \dfrac{\alpha}{1-\lambda(RF^{\varepsilon}-\beta)/\alpha} & \text{if } RF^{\varepsilon} < \beta \\ \alpha + \lambda(RF^{\varepsilon}-\beta) & \text{otherwise} \end{cases} \tag{5}$$

where λ and α represent reduced currents. β is the threshold after which the response of the cells becomes linear. Based on previous authors [16] we used $\lambda = 3$, $\alpha = 1$, $\beta = 0$. Such rectification is a common feature in retinal models [3]. It simulates static nonlinearities observed experimentally in the retina. Our aim with this formula is to imitate the response of a type of ganglion cells.

Based on the LN model, we propose the three binary descriptors depicted in figure 2. Each descriptor is constructed considering different components of ganglion cell response. The first component (D_C), just considers the response of the center of each receptive field. The second one (D_S), adds some information about the sign of the DoG. The last one (D_{ONOFF}), calculates the DoG of two population of cells, namely, ON and OFF cells. In the following each of them is explained in detail.

The center response The first component of our descriptors is defined considering the response of the center of the receptive field (RF). For all the cells in our pattern, we blur the center with the corresponding Gaussian kernel $K(x,y)$ obtained from equation 3, where the surround kernel G_{σ_C} is equal to 0. The

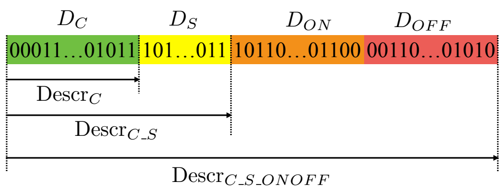


Fig. 2. Illustration of the construction of our descriptors from its components

activity in the center of the RF is calculated using equation 4. The purpose of this component is to mimick the original FREAK by performing the Gaussian smoothing as opposed to the approximation with the mean intensity calculation.

This binary component is constructed by calculating the difference in activity between all pairs of receptive fields.

$$\mathrm{D}_C(i,j) = \begin{cases} 1 & \text{if } N(RF_i) - N(RF_j) >= 0, \ \ \forall i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

In the following components of our proposal, the inhibitory effect of surround is taken into account.

The sign of the center-surround response This component takes into account the sign of the DoG centered in each of the 43 receptive fields of our model.

$$RF = sign(N(I * K(x,y))) \ . \tag{7}$$

As a result, the binary component is calculated as:

$$\mathrm{D}_S(i) = \begin{cases} 1 & \text{if } N(RF_i) >= 0, \ \ \forall i \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

In this way the contribution of adding some information about the inhibitory surround is evaluated.

The ON and OFF cell response Finally, the responses of both ON and OFF ganglion cells are considered. The activation of the ON cells is calculated from the formula 4, where ε is equal to $+1$. In a similar way, the activation of the OFF cells is calculated considering ε equal to -1.

The binary component is constructed comparing the activation between pairs of cells. For instance, for ON cells:

$$\mathrm{D}_{ON}(i,j) = \begin{cases} 1 & \text{if } N(RF_i) - N(RF_j) >= 0, \ \ \forall i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

The activity of the OFF cells $D_{OFF}(i,j)$ is encoded in an analogous way. In summary, this component, named $D_{ONOFF}(i,j)$, is constructed upon the concatenation of $D_{ON}(i,j)$ and $D_{OFF}(i,j)$.

However, depending on the images used as stimulus, the response of ON and OFF cells can be noisy. As a result the encoded information leds to less discriminative descriptors. Since our aim is to construct sparse descriptors, we implemented a variant of the $D_{ONOFF}(i,j)$ component. In an attempt to reduce noise, the activity of these cell types has been thresholded based on the neighborhood information [6]. For a given image, a pyramid of DoG is calculated corresponding to the 8 different cell sizes in our pattern. The average DoG is used to filter those pairs whose activity difference is above the threshold.

For instance, for the ON cells we consider:

$$\mathrm{D}_{ON_{Th}}(i,j) = \begin{cases} 1 & \text{if } N(RF_i) - N(RF_j) >= T, \ \forall i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where T is the average DoG. For the OFF cells, the activity response is thresholded in the same way. As a result, the variant of the ON and OFF cell response, named $D_{ONOFF_{Th}}(i,j)$, is composed of the thresholded ON cell pairs $D_{ON_{Th}}$ and the thresholded OFF cell pairs $D_{OFF_{Th}}$. Therefore, the corresponding descriptor is named $\mathrm{Descr}_{C_S_ONOFF_Th}$and is similar to third descriptor shown in figure 2, but considering the thresholded ON and OFF responses. All our descriptors have been tested on the scene classification task. The pipeline used in order to achieve this goal is described in the next section.

2.3 BOW approach for scene categorization

Scene categorization is accomplished using a bag-of-words (BOW) approach. The descriptors are densely extracted from the images using a grid at steps of 5 pixels. Each position of the grid is considered as a keypoint and the sampling grid is situated on top of it. For the original FREAK, the pattern size is 45x45 pixel. Since our sampling grid is slightly bigger because the surround is added to the original pattern, the size of our pattern is 60x60 pixel. As a result we obtain overlapping patches, which has been shown to be efficient for the image categorization task [11].

Regarding the descriptor size, for the original FREAK is 64 bytes because only selected pairs are considered, whereas our descriptors are larger since all the possible pairs are taken into account. We made all of them 512 bytes long, adding padding where necessary. All the descriptors are quantized into visual words by k-means, using as distance metric the Euclidean distance. Spatial pyramid histograms are used as kernels. After the training phase, the final classification is performed using a linear SVM.

3 Performance evaluation

We evaluated the performance of our descriptors on the 15 scene categories dataset [4], which is an extension to the 8 scene categories provided by [7]. As established in previous works [4], from each class 100 images are used as training and the rest as test. In total, 1500 images have been used for the training set and 2985 for the test set. All the tests have been done using 10 random splits. We used 100 randomly selected images from the training set to form the dictionary. In the table 1 we can see the mean accuracy of each approach using 600 visual words. The code for FREAK is available in openCV [2] and our descriptors have been implemented based on that code. The BOW approach is based on the VLFeat toolbox [12].

Table 1. Comparison of FREAK and our descriptors for image classification on the 15 scene categories dataset [4].

	FREAK	Descr$_{C}$	Descr$_{C_S}$	Descr$_{C_S_ONOFF}$	Descr$_{C_S_ONOFF_Th}$
mean	66.42%	67.93%	68.42%	70.37%	72.19%
std	±0.45	±0.70	±0.41	±0.68	±0.6

As a baseline we used the original FREAK, where selected pairs are used retaining the most informative and discriminative ones. Moreover, the more relevant pairs correspond to the outer region of the pattern, suggesting that first the periphery of an object is explored in an attempt to find interesting regions, mimicking the saccadic search. However, the gaussian filtering is approximated calculating the mean intensity inside each receptive field.

In our descriptors we considered all the possible pairs, since the ones selected in [1] are obtained after learning the best pairs from their training data. Experimentally we obtained better results when all the pairs are considered. As is shown in the table 1, all our descriptors are able to perform better than FREAK. The drawback is that the size of the descriptors is larger and there can be correlations between pairs. In the table 2 we show preliminary results obtained by reducing the dimensionality of the $Descr_{OnOff_{th}}$, using the same 10 random splits as in table 1.

Table 2. Effect of PCA dimensionality reduction: PCA applied to the thresholded response of ON and OFF cells Descr$_{C_S_ONOFF_Th}$

Eigenvectors	Descr$_{C_S_ONOFF_Th}$
64	72.11% ±0.61
128	72.46% ±0.75
256	72.78% ±0.70

In this table we can observe that eliminating the less discriminative pairs from the descriptor increases the performance. Best results are obtained when the size is reduced to 256 bytes. In comparison, our approach outperforms the original FREAK even when both methods use the same size of descriptors (i.e 64 bytes). In addition to this, in all our experiments the scale and orientation normalization is not used, since we are using a dense grid and not a keypoint detector as in the original idea.

The confusion matrix from one run of the $\mathrm{Descr}_{C_S_ONOFF_Th}$descriptor is shown in figure 3, where row names are true labels and column names are the predicted ones. The highest confusion happens between category pairs such as inside city/tall building, coast/open country, forest/mountain, bedroom/living room, industrial/store, which has been previously stated by other authors [4] [17].

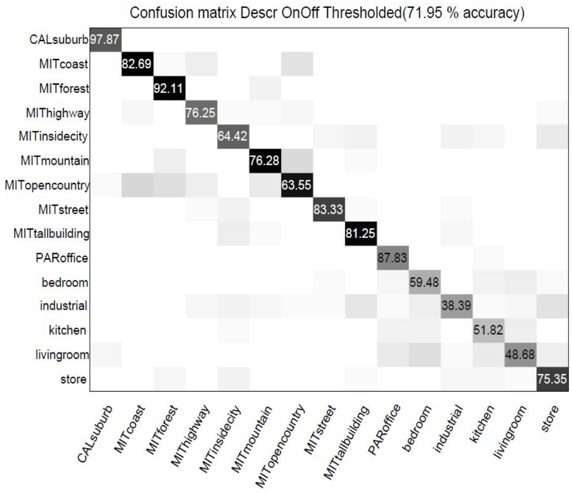


Fig. 3. Confusion matrix showing results obtained using our $\mathrm{Descr}_{C\_S\_ONOFF\_Th}$ descriptor on the 15 scene categories dataset [4].

A further analysis of the results is shown in the figure 4, where the mean accuracy results obtained with both the original FREAK and our descriptor $\mathrm{Descr}_{C_S_ONOFF_Th}$are plotted. In comparison our descriptor is able to classify better all the classes with the exception of two, namely, tallbuilding and living room. But for all the rest, our proposal outperforms the original FREAK. Overall, our third thresholded descriptor is able to achieve a high accuracy in recognizing natural scene categories, such as forest and coast. However, the results drop for most of the indoor scenes, as can be observed in the graph. There are other works related to scene classification that have reported the same issue. As explained in [9], the main two reasons could be, on the one hand, the lack of a large dataset of indoor scenes to train and test the approaches and, on the other hand, the difficulty in characterizing such scenes, which mainly requires a combination of both local and global image information. Interestingly, store and office images are classified much better with our descriptor than with FREAK, which suggests that our approach is able to better represent the properties of those type of images.

In this paper, we have proposed to extend the original FREAK in the following way. Our first descriptor, Descr_C, blurs the center of each receptive field with a Gaussian kernel. Since we used the same kernel size as the original FREAK, the results obtained with this modification are similiar in both cases. Our sec-

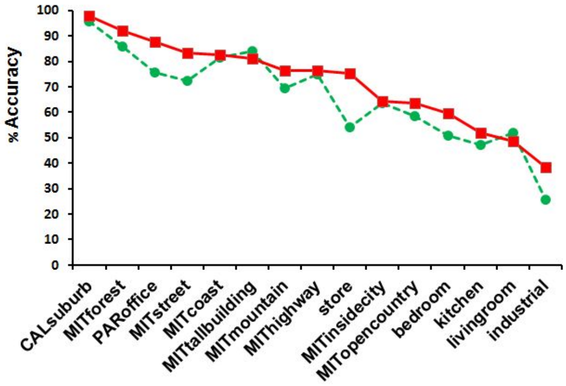


Fig. 4. Comparison of the mean accuracy percentage between FREAK (green curve) and our $\text{Descr}_{C\_S\_ONOFF\_Th}$ descriptor (red dashed curve) for each image class.

ond descriptor, Descr_{C_S}, adds some information about the sign of the DoG. We tested the contribution of using both the center response D_C and the DoG sign D_S, improving the performance slightly. Our third descriptor, $\text{Descr}_{C\_S\_ONOFF}$, implements the DoG considering two population of cells, namely, ON and OFF cells. In comparison with FREAK (66.42%) the results obtained with our descriptor are better (70.37%). Moreover, when the cell activiy is thresholded the classification accuracy is improved. If instead of using the D_{ONOFF} as our third component, we use the tresholded response D_{ONOFF_Th} the accuracy obtained is 72.19%. In addition to this, PCA dimensionality reduction applied to this descriptor further improves the results, achieving a 72.78% of correct classifications.

4 Conclusions

The goal of this work was to implement a bio-inspired descriptor, mimicking some functionalities of the visual system. From biology, it is well known that the retina extracts details from images using a Difference of Gaussians (DoG) of different sizes and encodes such differences with action potentials. We have presented a set of modifications to FREAK which are more biologically inspired. As a conclusion it seems that difference of gaussians calculated inside each receptive field, as is done by the visual system, extracts useful information for scene classification task. In the future, other low-level processing performed by the retina could be considered. In relation with this, other organization of the cells can also be tested, since as stated by Alahi et al. [1], changing the size of the receptive fields and their overlap increases the performance. Finally, the dimensionality of our descriptors can be reduced learning the most significant pairs in our model. Potentially, retaining the most significant pairs will further improve the classification results.

Acknowledgement

We thank M. San Biagio for his support in the image classification algorithm. This research received financial support from the 7th Framework Programme for Research of the European Commission, under Grant agreement num 600847: RENVISION project of the Future and Emerging Technologies (FET) programme Neuro-bio-inspired systems (NBIS) FET-Proactive Initiative

References

1. A. Alahi, R. Ortiz, and P. Vandergheynst. FREAK: Fast Retina Keypoint. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 510–517, (2012).
2. G. Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools*, (2000).
3. E. J. Chichilnisky. A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, 12(2):199–213, (2001).
4. 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. IEEE Computer Society, (2006).
5. S. Leutenegger, M. Chli, and R. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *ICCV'11*, pages 2548–2555, (2011).
6. X. Meng, Z. Wang, and L. Wu. Building global image features for scene recognition. *Pattern Recogn.*, 45(1):373–380, (2012).
7. A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, (2001).
8. A. Oliva and A. Torralba. Building the gist of a scene: the role of global image features in recognition. In *Progress in Brain Research*, volume 155, pages 23 –36, (2006).
9. A. Quattoni and A. Torralba. Recognizing indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 413 – 420, (2009).
10. E. Tola, V. Lepetit, and P. Fua. DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830, (2010).
11. T. Tuytelaars. Dense interest points. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 2281–2288, (2010).
12. A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, (2008).
13. N.-S. Vu, T. P. Nguyen, and C. Garcia. Improving texture categorization with biologically inspired filtering. *Image and Vision Computing*, 32:424–436, (2013).
14. J. Wang, X. Wang, X. Yang, and A. Zhao. Cs-freak: An improved binary descriptor. In *Advances in Image and Graphics Technologies*, pages 129–136. Springer, (2014).
15. C. Whiten, R. Laganiere, and G.A. Bilodeau. Efficient action recognition with mofreak. In *Proceedings of the 2013 International Conference on Computer and Robot Vision*, pages 319–325. IEEE Computer Society, (2013).
16. A. Wohrer. *Model and large-scale simulator of a biological retina with contrast gain control*. PhD thesis, University of Nice Sophia-Antipolis, (2008).
17. J. Wu and J. M. Rehg. Centrist: A visual descriptor for scene categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1489–1501, (2011).