Scale Invariant Dissociated Dipoles

Marco P. Vanossi[1] and Julian Stöttinger[2]

[1] University of Campinas, São Paulo, Brazil

[2] Institute for computer-aided automation,
Vienna University of Technology, Austria

Abstract
In this paper we investigate the use of dissociated information to represent image structures in a scale invariant way. Our contribution consists in an extension to the dissociated dipoles based descriptor that has proven to be as robust to image transformations as SIFT, while containing 6 times less data. We demonstrate that our extension has better performance when scaling of images takes place. Further, we show the benefits of the more stable color interest points for both feature localization and scale selection. We are able to demonstrate that colored scale invariant dipoles have comparable or better matching scores than SIFT in an exact nearest neighbor matching scheme.

1. Introduction

Many current object recognition systems use as basic components local features. A salient local feature describes an image pattern which is different to its immediate neighborhood. Most commonly, the image property considered is intensity. Local features are usually extracted at interest points (e.g. [5, 11]). The description of these features can be done e.g with histograms to represent different characteristics of the patch. Differential descriptors are employing a set of image derivatives computed up to a given order to approximate a point neighborhood. Spatial-frequency based descriptors describe the frequency distribution of an image patch. An evaluation of such descriptors was done by Mikolajczyk et al. [15] and concluded that SIFT descriptors performed best. In this paper we investigate the use of non-local differential operators to describe the interest points. Such operators have been introduced by Ballas et al. [2] motivated by the fact that comparing small regions across large distances is more robust to common image transformations than performing local evaluations, suggesting they may provide a useful vocabulary for image encoding.

The Dissociated Dipole or Sticks operator has then been successfully used in gray-scale converted images in a using different fixed scales by Joly et al. [9]. In the evaluation campaign ImagEval[1] of content-based image retrieval techniques it showed to provide promising performance. In such work, a multi-resolution scheme is applied to sustain small scale changes. We propose an improvement to the technique by adding more robust scale selection. We also use color information for both detecting and selecting the characteristic scale of the features.

This paper is organized as follows. The following Section 2. briefly reviews local features; Section 3. presents how dipoles are extracted and how to derive scale invariance; Section 4. provides experimental validation while Section 5. concludes.

[1] `imageval.org`

2. Local Features

Local features can be points, but also edgels or small image patches. Typically, some measurements are taken from a region centered on a local feature and converted into descriptors.
Initially proposed detectors extracted features at a single scale. To deal with scale changes scale-invariant methods have been introduced. These automatically determine both the location and scale of the local features. The existing approaches mainly differ in the differential expression used to build the scale-space representation. A normalized LoG function was applied in [12] to build a scale space representation and search for 3D maxima. The scale-space representation is constructed by smoothing the high resolution image with derivatives of Gaussian kernels of increasing size. Automatic scale selection is performed by selecting local maxima in scale-space. In [14] a scale invariant corner detector, Harris-Laplace, and a scale-invariant blob detector, Hessian-Laplace, were evaluated. In these methods, position and scale are iteratively updated until convergence. These detectors are similar to the DoG approach [13], which localizes points at local scale-space maxima of the difference-of-Gaussian. To cope with non-uniform scaling and skew, affine invariant detectors have also been proposed, e.g. [18]. An approach which successfully deals with scale changes in a different way are MSER [4] which are extracted with a watershed like segmentation algorithm.

Many different techniques for describing local image regions have been developed. Techniques that use histograms to represent different characteristics of appearance or shape are the tool of choice for robust description. The most popular descriptor is the one proposed by Lowe [13]: a scale invariant feature transform (SIFT), which combines a scale invariant region detector and a descriptor based on the gradient distribution in the detected regions. The descriptor is represented by a 3D histogram of gradient locations and orientations. The contribution to the location and orientation bins is weighted by the gradient magnitude. The quantization of gradient locations and orientations makes the descriptor robust to small geometric distortions and small errors in the region detection. Geometric histogram [1] and shape context [3] implement the same idea and are very similar to the SIFT descriptor. Both methods compute a 3D histogram of location and orientation for edge points where all the edge points have equal contribution in the histogram. These descriptors were successfully used, for example, for shape recognition of drawings for which edges are reliable features. Yan Ke et al. [10] proposed to robustify the SIFT descriptors. Their descriptors encode the salient aspects of the image gradient in the feature point's neighborhood; however, instead of using SIFT's smoothed weighted histograms, they apply Principal Components Analysis (PCA) to the normalized gradient patch resulting in a more compact representation.

Another approach are "spin images" [8] introduced for 3D object recognition in the context of range data. Their representation is a histogram of the relative positions in the neighborhood of a 3D interest point. The two dimensions of the histogram are distance from the center point and the intensity value. Zabih and Woodfill [22] have developed an approach robust to illumination changes. It relies on histograms of ordering and reciprocal relations between pixel intensities which are more robust than raw pixel intensities. This descriptor is suitable for texture representation but a large number of dimensions is required to build a reliable descriptor [17].

The Fourier transform decomposes the image content into the basis functions. However, in this representation the spatial relations between points are not explicit and the basis functions are infinite, therefore difficult to adapt to a local approach. The Gabor transform [6] overcomes these problems, but a large number of Gabor filters is required to capture small changes in frequency and orientation. Gabor filters and wavelets [21] are frequently explored in the context of texture classification.

3. Scale Invariant Colored Dipoles

Like a simple edge detector, a Dissociated Dipole is a differential operator consisting of an excitatory and an inhibitory lobe (see Fig. 1 for an schematic representation), and may be used at any orientation or scale. However, unlike a conventional edge detector, it allows an arbitrary separation between these two lobes, removing the correlation of inter-lobe distance and lobe size. Therefore, the dipole filters take advantage of having a flat variation at their center, providing them a better robustness to localization errors, as shown by Fig. 2.

At an early stage, many image retrieval scenarios use interest point detection to find regions in which descriptors are calculated. We aim to find the most robust locations by using spherical color spaces and a stable scale selection by projecting this color information. First the scale space of the Harris function is built iteratively

$$E(x, y, s) = (x, y)M\left(x, y, t^s\sigma_D, \tfrac{t^s}{3}\sigma_I\right)\begin{pmatrix} x \\ y \end{pmatrix} \tag{1}$$

under varying σ_D and σ_I. As shown in several experiments [14], the relation $\sigma_D = 3\sigma_I$ performs best. We use scale steps $s = 1, 2, \ldots$ determining the iterations of the algorithm (typically between 8 and 20) with a factor t from 1.2 to $\sqrt{2}$. The amount of scale change is chosen by the need for preciseness of the corner location. Also, as an extension of the Harris detector to color, the second moment matrix is defined as proposed in [16]

$$M = \sigma_D^2 G_{\sigma_I} \otimes \begin{bmatrix} R_x^2 + G_x^2 + B_x^2 & R_xR_y + G_xG_y + B_xB_y \\ R_xR_y + G_xG_y + B_xB_y & R_y^2 + G_y^2 + B_y^2 \end{bmatrix} \tag{2}$$

where $\otimes$ indicates convolution and the subscripts x and y indicate Gaussian derivatives at scale σ_D in these directions. σ_I is the integration scale. We follow the extension to arbitrary color spaces [20] and use the gradients of the transformed color space instead of the original RGB values. We adopted the *HSI* color space. To choose the characteristic scale the LoG function on the projected image is used: We apply PCA to the transformed color gradients, ensuring a trade-off between favoring rare colors and retaining information on similar colors. When both the Harris Energy and the LoG are extrema a possible region is found [19]. Given a set of color interest points computed as above, the 20-dimensional dipole descriptor is computed as follows.

To achieve rotational invariance, each interest point is assigned a orientation following [13]: An orientation histogram is formed from the gradient orientations within a region around $\boldsymbol{P}$ and the dominant direction θ_0 is then estimated by the highest peak in this histogram. Any other local peaks that is within 80% of the highest peak is used to create a new interest point with the corresponding orientation. For better accuracy, the position of the peaks are interpolated by a parabolic fit on 3 histogram values.

Each dipole consists of a pair of Gaussian lobes, with standard deviation σ and a spatial separation of δ, and is computed as a difference of two values in a simple Gaussian Scale-Space. In order to be fully invariant to scale changes, the spatial separation δ_1 is set to the characteristic scale of the interest point and the size of the Gaussian window σ_1 is defined by $\frac{\delta_1}{2}$. The descriptor is composed of two

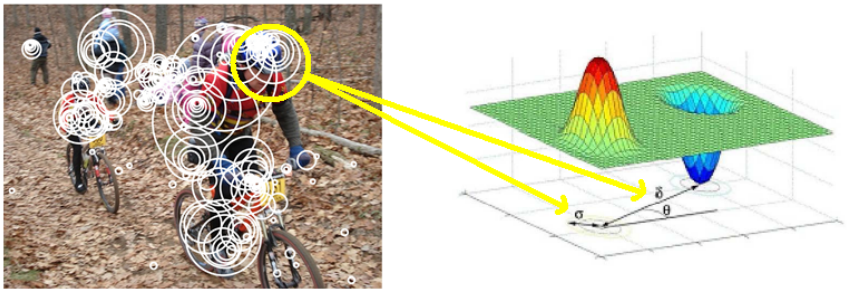


Figure 1. Using interest point scale information to derive scale invariant dipoles. The prototypical dipole scheme from [2] is modified to cope with image resizes.

sub-vectors $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$. $\boldsymbol{F_1}$ is composed of 8 dipoles and has scale σ_1, while $\boldsymbol{F_2}$ has 12 dipoles with half scale.

First order dipoles: Let $\boldsymbol{G}$ be the vector composed of the values of $\boldsymbol{L}(x,y,\sigma_1)$ in 12 dimensions at a distance δ_1 around the interest point $\boldsymbol{P}$, it's i-th component g_i being defined as $g_i = \boldsymbol{L}(x_i, y_i, \sigma_1)$ with $x_i = x_c + \delta_1 \cdot \cos(\theta_i)$, $y_i = y_c + \delta_1 \cdot \sin(\theta_i)$. where (x_c, y_c) is the position of the interest point $\boldsymbol{P}$ and

$$\theta_i = \theta_0 + (i-1) \cdot \frac{2\pi}{12} \tag{3}$$

First order vector $\boldsymbol{F_1}$ is then obtained by forming $D_1 = 8$ dipoles from the components of $\boldsymbol{G}$ according to the linear relation $\boldsymbol{F_1} = \boldsymbol{A} \cdot \boldsymbol{G}$ where

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{4}$$

Second order dipoles: $\boldsymbol{F_2}$ is composed of $D_2 = 12$ dipoles of parameters $\sigma_2 = \frac{\sigma_1}{2}$ and $\delta_2 = \frac{\delta_1}{2}$. They are computed along 12 orientations at a distance δ_1 around the interest point $\boldsymbol{P}$ according to:

$$\boldsymbol{F_2} = \begin{pmatrix} f_1^2 \\ .. \\ f_i^2 \\ .. \\ {f_{D_2}}^2 \end{pmatrix}, \text{where } f_i^2 = \boldsymbol{L}(x_i', y_i', \sigma_2) - \boldsymbol{L}(x_i'', y_i'', \sigma_2) \text{ and } \begin{bmatrix} x_i' = x_c + (\delta_1 + \delta_2).\cos(\theta_i) \\ y_i' = y_c + (\delta_1 + \delta_2).\sin(\theta_i) \\ x_i'' = x_c + (\delta_1 - \delta_2).\cos(\theta_i) \\ y_i'' = y_c + (\delta_1 - \delta_2).\sin(\theta_i) \\ \theta_i = \theta_0 + (i-1).\frac{2\pi}{12} \end{bmatrix} \tag{5}$$

In order to be invariant to affine luminance transformations (of the kind $a\boldsymbol{I}(x,y) + b$), the two sub-vectors $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$ are normalized on a sphere by dividing them by their L_2-norms $||\boldsymbol{F_1}||_2$ and $||\boldsymbol{F_2}||_2$. Invariance to negative and flip can also be obtained by the same processes applied in [9].

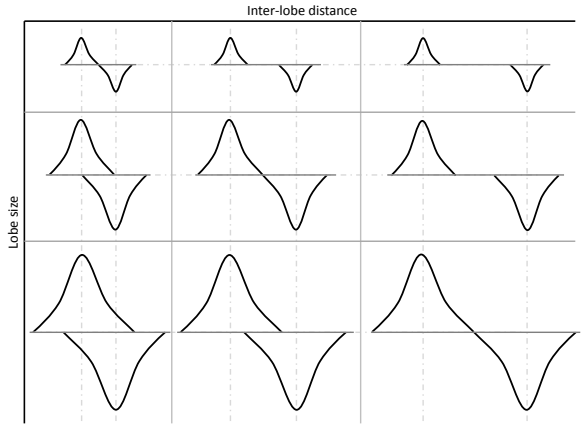
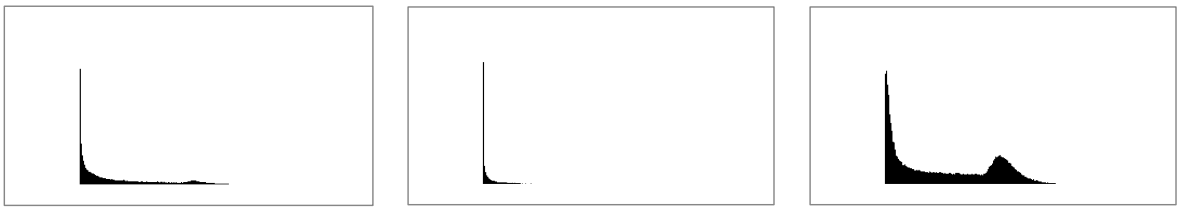


Figure 2. De-coupling of the parameters of inter-lobe distance and lobe size [2]. Conventional edge detectors confound these two attributes and, effectively, they use only the diagonal elements of the space defined by these parameters.

4. Experiments

In this section we show that even though our feature has 6 times less information, its representation is much more discriminative as the distribution of the data along while its dimensions is populated more densely. We perform exact matching experiments to show the efficiency of the technique. In the following, we analyze the population and density of the feature space on a subset of the well known ALOI database[2] [7]. We find significant differences in the distribution of the values in the proposed Dipoles and SIFT. In Section 4.2. we show that the we gain comparable or slightly improved results in image matching using 6 times less data than SIFT.

4.1. Features Distribution



(a) SIFT coordinate 0 (b) SIFT coordinate 10 (c) SIFT coordinate 48

Figure 3. Histograms of SIFT vectors values along several dimensions on the ALOI subset

Figure 3 shows the distribution of the coordinates of SIFT vectors on three chosen directions. These figures were obtained from SIFT descriptors of images from the ALOI subset where the illumination direction changes. Extracting features from the ImagEval benchmark 1[3] similar results are obtained. Fig. 3(a) is almost representative of all the 128 histograms. Just a few ones are different (see Fig. 3(c)). These different distributions are a consequence of the SIFT vectors construction: Each dimension is a bin where local gradients of a given direction are accumulated. This direction is measured relatively to the major direction of the SIFT descriptor. Thus, local bins which represent gradients of the same

[2] staff.science.uva.nl/~aloi/

[3] imageval.org

direction as the major one are naturally the largest. Histogram of 3(c) corresponds to dimension 48 which accumulated gradient in a direction equal to the major one. The consequence is that coordinates of the 128 dimensional descriptors which are very low are much more common than those with high values. On the other hand, Figure 4 shows that the Dipoles distribution is much denser, resulting in a more discriminative representation.

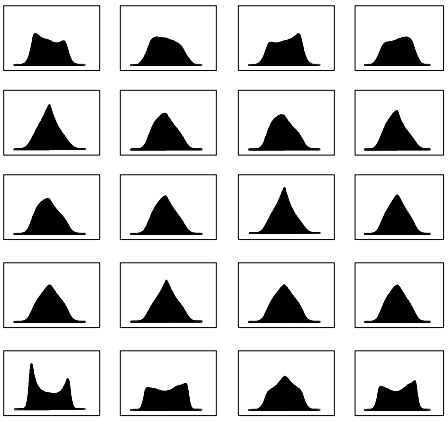
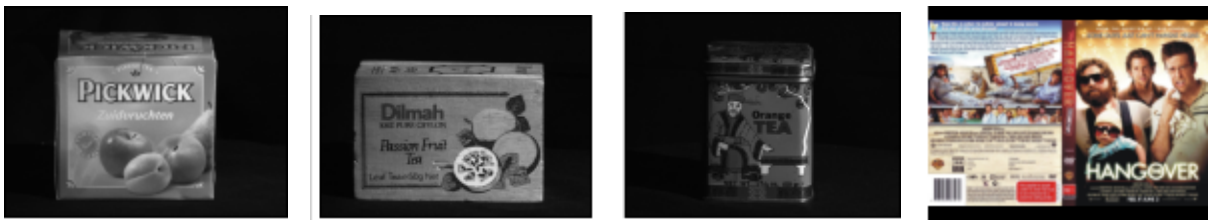


Figure 4. Distribution of the dimensions of the Dipoles dataset is dense.

4.2. Matching



(a) ALOI object 18 (b) ALOI object 110 (c) ALOI object 78 (d) DVD cover

Figure 5. Example images for the matching experiments.

	Luminance based interest points	
img	SIFT	DIPOLE
13	0,8	**0,82**
18	0,69	**0,75**
26	0,81	**0,82**
46	0,83	**0,85**
78	0,48	**0,49**
101	0,79	**0,80**
110	0,68	**0,71**
216	0,81	**0,84**
219	**0,71**	0,68
264	0,77	**0,80**
mean	0,716	0,723

Table 1. Matching performance on ALOI images

To validate the improved description based on color localization and more compact description of local surrounding, we carry out matching experiments on the ALOI database and DVD covers (see

Fig. 5). As shown in Tbl. 1, we gain improved results on nearest neighbor matching on luminance based interest points compared to SIFT.

Using a more stable scale selection, the gain in matching performance is significantly boosted compared to SIFT. This is observed for the matching of DVD covers. We match two DVD covers one being scaled by 50%. DVD covers typically consist of colorful patterns and therefore we can gain improved matching results using color interest point detection and Dipoles. The proposed method uses 6 times less data than SIFT, but distributes the described information better throughout the feature space.

	HSI color points	
img	SIFT	DIPOLE
cooking mama	0.35	**0.46**
ratatouille	0.14	**0.16**
happy feet	0.12	**0.29**
hangover	0.31	**0.38**
shrek	**0.24**	0.23
cars	**0.17**	0.16
forest gump	0,175	**0.23**
kung fu panda	**0.33**	0.3
batman	**0.26**	0.24
indiana jones	0.205	**0.23**
mean	0.23	**0.27**

Table 2. Matching performance SIFT vs DIPOLE on 50% scaled DVD covers

5. Conclusion and future work

We introduced in this paper an approach that more compact and discriminatingly encodes the content of images while giving state of the art matching results. We adapted the descriptor based on dissociated dipoles to be used for scale invariant local description. Experiments showed that the descriptor dimensions are better distributed in the feature space than SIFT, providing a robust description of visual data with 6 times less information than the state of the art. Future work will include the incorporation of color information into the description phase and the contribution of the dipoles to different image indexing techniques.

Acknowledgment

This work was partly supported by the Austrian Research Promotion Agency (FFG) and the CogVis[4] FFG or CogVis Ltd. are not liable for any use that may be made of the information contained herein.

References

[1] A. P. Ashbrook, N. A. Thacker, P. I. Rockett, and C. I. Brown. Robust recognition of scaled shapes using pairwise geometric histograms. In *BMVC*, pages 503–512, 1995.

[2] B. J. Balas and P. Sinha. Dissociated dipoles: Image representation via non-local comparisons. In *CBCL*. MIT Press, 2003.

[3] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24(4):509–522, 2002.

[4] http://www.cogvis.at/

[4] Maximally Stable Extremal, J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from. In *BMVC*, pages 384–393, 2002.

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, volume 2, pages 264–271, June 2003.

[6] D. Gabor. Theory of communication. *J. IEE*, 3(93):429–457, 1946.

[7] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders. The amsterdam library of object images. *IJCV*, 61(1):103–112, 2005.

[8] Andrew E. Johnson and Martial Hebert. Recognizing objects by matching oriented points. In *CVPR*, pages 684–689, 1996.

[9] A. Joly. New local descriptors based on dissociated dipoles. In *CIVR*, pages 573–580, 2007.

[10] Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *CVPR*, pages 506–513, 2004.

[11] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC*, pages 759–768, 2003.

[12] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30:79–116, 1998.

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

[14] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004.

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

[16] P. Montesinos, V. Gouet, and R. Deriche. Differential invariants for color images. In *ICPR*, 1998.

[17] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 24(7):971–987, 2002.

[18] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In *ECCV*, pages 414–431, 2002.

[19] J. Stottinger, A. Hanbury, T. Gevers, and N. Sebe. Lonely but attractive: Sparse color salient points for object retrieval and categorization. In *CVPRW*, pages 1–8, 2009.

[20] J.t van de Weijer, T. Gevers, and J. M. Geusebroek. Edge and corner detection by photometric quasi-invariants. *PAMI*, 27(4):625–630, 2005.

[21] J.K.M. Vetterli. Wavelets and subband coding. *Prentice Hall*, 1995.

[22] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *ECCV*, pages 151–158, 1994.