A comparison of dense region detectors for image search and fine-grained classification

Ahmet Iscen, Giorgos Tolias, Philippe-Henri Gosselin, Hervé Jégou

A comparison of dense region detectors for image search and fine-grained classification

Ahmet Iscen*, Giorgos Tolias, Philippe-Henri Gosselin and Hervé Jégou

Abstract—We consider a pipeline for image classification or search based on coding approaches like Bag of Words or Fisher vectors. In this context, the most common approach is to extract the image patches regularly in a dense manner on several scales. This paper proposes and evaluates alternative choices to extract patches densely. Beyond simple strategies derived from regular interest region detectors, we propose approaches based on super-pixels, edges, and a bank of Zernike filters used as detectors.

The different approaches are evaluated on recent image retrieval and fine-grain classification benchmarks. Our results show that the regular dense detector is outperformed by other methods in most situations, leading us to improve the state of the art in comparable setups on standard retrieval and fined-grain benchmarks. As a byproduct of our study, we show that existing methods for blob and super-pixel extraction achieve high accuracy if the patches are extracted along the edges and not around the detected regions.

Index Terms—dense keypoints, image retrieval, fine-grained classification, Zernike polynomials

I. INTRODUCTION

LOCAL image description is a popular research topic in computer vision, as it is involved in many applications such as image classification and particular object detection. Extracting local descriptors from an image consists of two steps. The *detection* step selects regions of interest, which are normalized into fixed-size patches. The *description* step produces a vector representation for each of the detected patches. The SIFT descriptor [29] and its RootSIFT extension [3] have been shown to perform very well for most applications. Many descriptors have been introduced in the last years to improve the description speed or descriptor compactness, such as SURF [6], CHOG [10] or BRIEF [9]. The matching accuracy is improved by learning the descriptor design [58][59][49]. In this paper, we are solely interested in the detection stage and therefore adopt the gold-standard SIFT and RootSIFT.

The early works in this line of research have focused on the detection of sparse interest points and regions, typically producing a few thousand descriptors per image. These approaches aim at extracting distinctive and repeatable image parts, such as blobs and corners [29][35][31][6], offering covariance properties: the same regions should be detected under some geometrical transformations. From a historical perspective, the choice of sparse representations was arguably motivated by the lack computational and memory resources. Although such methods perform well and are still widely used for image matching, they are not competitive in other application scenarios such as image classification.

Local feature detection has recently shifted towards denser techniques. Dense sampling is an easy way to provide a large number of patches and a better coverage of the objects of interest. Fei-Fei and Perona [15] were the first to show that dense patches leads to better classification accuracy. Nowak *et al.* [38] argue that the key parameter for classification is the number of extracted patches. Similar conclusions hold for other tasks, like fine-grained classification [19] or action recognition in videos [56][22]. Recent works [61][12][51][52] also evidence that methods for image and particular object retrieval, which traditionally rely on sparse regions typically extracted with the Hessian-Affine detector [34], are improved when using a larger set of descriptors.

Nevertheless, regular dense sampling has serious limitations. Uniform sampling of patches ignores the image structure and extracts many uniform and uninformative patches. Additionally, the position of the features is less or not repeatable. This prevents the image engine from employing a spatial verification method, such as RANSAC [17][44], which typically filters out many outliers by enforcing the spatial consistency of the detected regions.

In this paper, our goal is to develop and evaluate dense detection strategies for image retrieval and fine-grained image classification. Our motivation is similar to that of Tuytelaars when she introduced "dense interest points" [53]: we consider solutions in between localized sparse interest points and dense strategies, in order to produce a large number of localized regions. We depart from using traditional evaluation metrics for detector evaluation, which do not reflect the final goal. For instance, the repeatability score reflects the effectiveness in detecting inliers, but is not directly related to the determination of class membership. Instead, we evaluate the performance with the metrics employed for the target application scenario: mean average precision for image retrieval and accuracy for fine-grained classification. We stress that better localized dense patches (see Figure 1) are more important for these tasks than in traditional image classification: the objects are more repeatable and distinguishing between two classes often rely on tiny details that suffer from being loosely localized.

We make the following contributions in this context of dense detection for image retrieval and fine-grained classification:

1) We propose strategies derived from standard interest point detectors (Harris, Hessian and DoG) to extract patches densely. In particular, we modify the detection process by relaxing the standard local maxima criterion so that it focuses on edges and not only corners. This, jointly with the optimization of the scaling factor, is shown to be a key to achieve higher performance.

2) We depart from the typical choice of fitting an ellipse to describe a region of interest extracted by blob detectors (MSER or super-pixels). Instead, we sample patches at several scales along the region's borders. This increases the performance significantly when considering a large number of patches per image. We also show that sampling on the edges produced by a state-of-the-art edge detector [13] offers competitive performance.

3) Finally, we propose two novel response filters to select the patch locations. First, we propose to use a bank of Zernike polynomials [60] as detectors. These filters have been proposed to construct descriptors [48], but to our knowledge not as detectors. Our second strategy is descriptor-oriented: the response for each pixel is simply the norm of local descriptor associated with the patch centered at this location. These two new approaches appear to perform best in most cases of our experiments.

This paper is organized as follows. Section II discusses previous work related to the purpose of our study. Section III describes existing approaches which are part of our evaluation, while in section IV we present modified schemes of existing detectors and the use of pseudo-Zernike polynomials for dense feature detection. Finally in section V we present the outcome of this study on 2 datasets on image retrieval and 4 datasets on fine-grained classification.

II. RELATED WORK

A great deal of works has focused on the detection of local interest points [29][6][34][31]. They are rather designed to be appropriate for image matching applications. The standard evaluation metrics for such detectors [35], *e.g.* repeatability, are designed to reflect sufficiency for matching applications. However, this could be far from the final application tasks we focus in this paper, *i.e.* fine-grained classification.

Typically, local interest points are tuned to detect image structures, such as corners and blobs, highly distinctive and repeatable. One of the exceptions is the work of Mikolajczyk *et al.* [33] where they focus on edges to represent objects. Similarly, in this study we argue that such a choice is beneficial for image retrieval and fine-grained classification.

Tuytelaars's work on dense interest points [53] has a motivation close to ours. Initially, a dense grid of patches is considered, and then for each point, local refinement of its location and scale parameters is performed. In particular, the point of maximum *interestingness* is selected within a bounded area. This choice suffers from quantization artifacts, since a true local maximum might not exist in this search area. Apparently, a large number of patches are located on smooth and uninformative regions.

A previous survey closely related to this study is the work of Nowak *et al.* [38]. They are also not interested in the repeatability of patches, but directly measure classification accuracy. According to them, detectors designed to obtain high repeatability perform the same as randomly selected patches. They observe that performance is, up to some extent, an increasing function of the number of points per image. Inspired by their finding, we conduct a similar evaluation and try to study different ways to control the average number of patches. Similarly, Avrithis and Rapantzikos [4] do not restrict comparison to repeatability, but further compared image retrieval performance, while considering the average number of points as a crucial parameter.

One of the parameters we consider in our study is the size of the measurement region for local descriptor extraction. This has been given particular attention by the recent work of Simonyan *et al.* [49], who shows that introducing a scaling factor increases retrieval performance. Similarly, the improved Hessian-Affine detector by Perdoch *et al.* [41] uses a larger measurement region to improve the performance.

In regular sampling local descriptors are agnostically extracted on a dense grid, and as a consequence many uninformative descriptors are derived from smooth regions. A significant improvement is achieved by filtering out descriptors with low ℓ_2-norm [19]. We combine this approach with examined detectors and study its effect on different setups.

Convolutional neural networks are very effective for image classification [27][14] and detection [18]. However, for the tasks addressed in this paper, namely fine-grained classification, image and particular object, this is not (yet) the case. Fisher vectors based on a dense representation and employing medium-sized codebooks and spatial coding achieved much better results in the fine-grained challenge [19]. For image retrieval, recent work demonstrates that CNNs [5] achieve excellent performance for aggressive operating points for compact representations. However, their best reported performance is lower than that of the state-of-the-art Fisher vectors, and outperformed by a large margin by state-of-the-art methods like the selective match kernels [51].
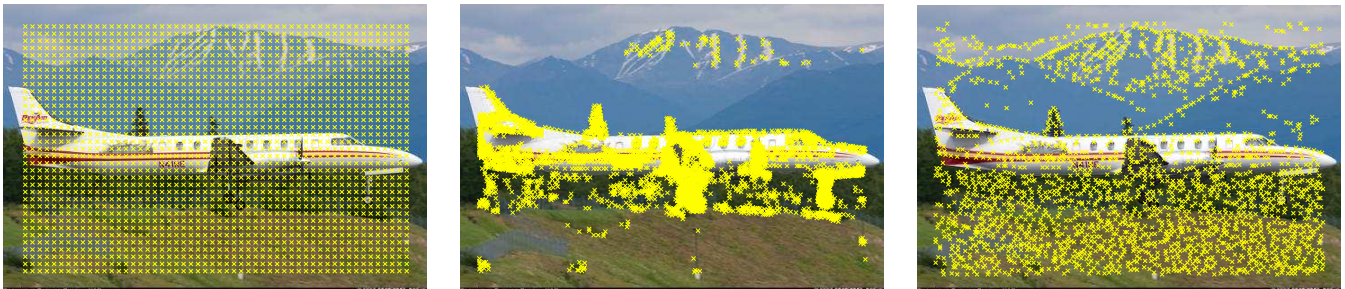

Fig. 1. Locations of patches detected by regular dense sampling (left), by Zernike detector (middle), and by dense ℓ_2-norm based detector (right). We visualize the patches only for the first scale.
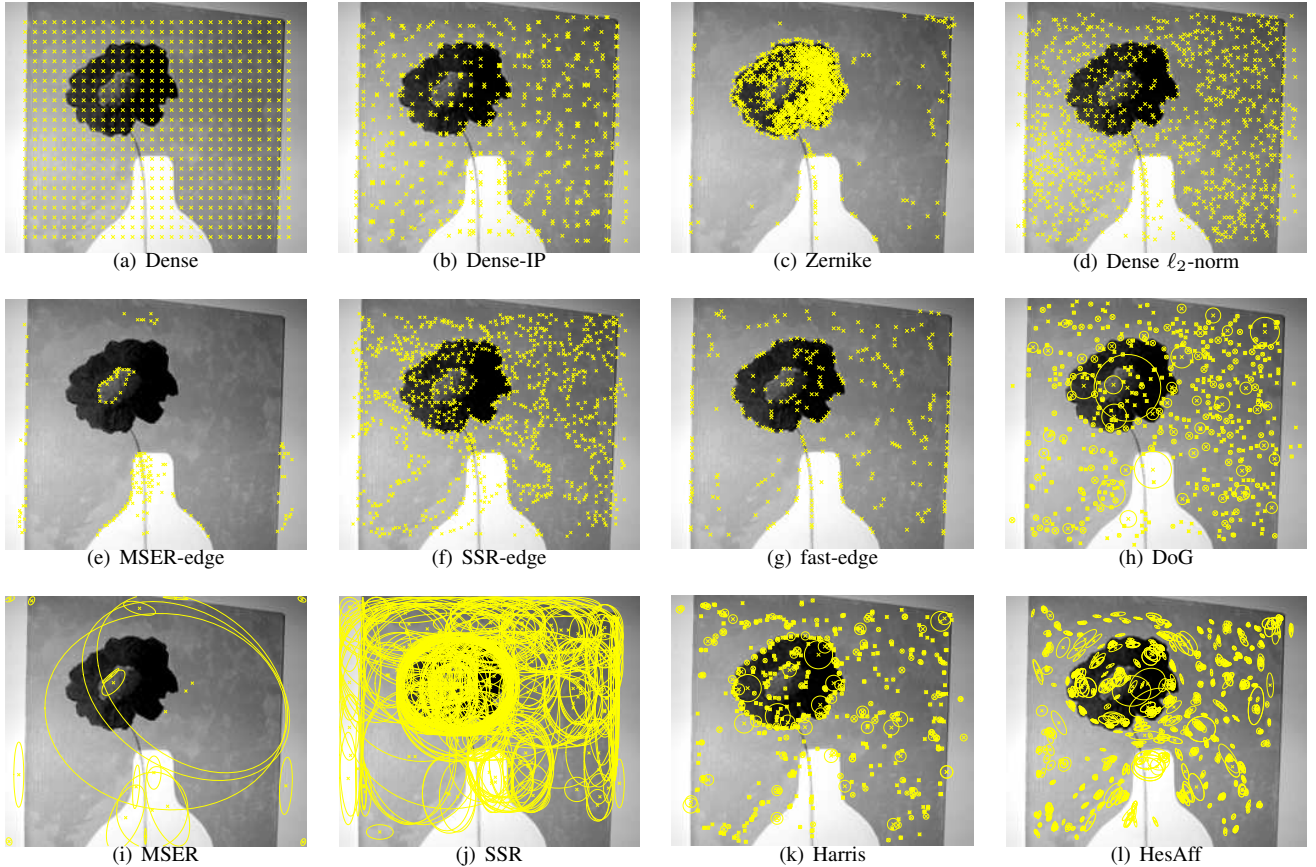
Fig. 2. Visualization of regions detected by each method. For methods with predefined scales, we visualize the center of points detected on the first scale (a)-(g), while for the ones that perform scale selection we draw the corresponding ellipses or circles (h)-(l).

III. BACKGROUND

In this section we briefly describe existing methods for interest point detection or dense patch sampling, that are part of our study. In all those methods, a local descriptor is extracted from each region of interest and an image is represented by a set of such descriptors. We visualize detected features for all the detectors examined or proposed in this work on Figure 2.

A. Interest points

Harris-Laplace detector. Harris detector localizes corners, based on the fact that gradient values will change in multiple directions around a corner. It uses a scale adapted version of the second moment matrix, known as Harris matrix [20]:

$$M = \sigma_D^2 \cdot g(\sigma_I) * \begin{bmatrix} L_x^2(x,\sigma_D) & L_xL_y(x,\sigma_D) \\ L_xL_y(x,\sigma_D) & L_y^2(x,\sigma_D) \end{bmatrix}, \quad (1)$$

where σ_D is the differentiation scale, σ_I is the integration scale and L_z is the derivative computed in z direction. Differentiation scale σ_D is used to compute the local derivatives with Gaussian kernels, and a Gaussian window with a size σ_I is used to smooth and average the neighborhood around the point. The eigenvalues of this matrix represent the gradient changes in two directions. Consequently, if one eigenvalue is large while the other is small, there exists an edge, whereas

if both eigenvalues are large, there exists a corner. The *interestingness* of a point is captured by the *cornerness* function, defined as $cornerness = \det(M) - \alpha\text{trace}^2(M)$, where α is usually set to 0.05.

Extending Harris detector to be scale invariant, Mikolajczyk *et al.* [34] use Laplacian-of-Gaussian (LoG) response and detect local extrema over multiple scales to perform scale selection [28]. This is the well known Harris-Laplace detector. Only points with cornerness value higher than threshold τ are retained and final point locations are chosen by a local maxima search procedure.

Hessian-Affine detector [35], similar to Harris detector, detects image locations that have large derivatives in both directions. Point locations are selected as local maxima of the Hessian matrix determinant, which now constitutes the interestingness measure. The Hessian matrix is defined as

$$H = \begin{bmatrix} L_{xx}(x,\sigma_D) & L_{xy}(x,\sigma_D) \\ L_{xy}(x,\sigma_D) & L_{yy}(x,\sigma_D) \end{bmatrix}, \quad (2)$$

where L_{zz} is the second order partial derivative. All points with interestingness below threshold τ are discarded. Although Hessian and Harris detectors are quite similar, the detected points may be slightly different. In particular, unlike Harris, Hessian detector tends to select locations with texture variations, in addition to corners.

The affine shape of the point is estimated by the eigenvalues of the second moment matrix M. An iterative procedure modifies the point's location, scale and shape until the estimated affine transform is able to map the detected region into one that has equal eigenvalues of its second moment matrix.

MSER *Maximally Stable Extremal Regions* were introduced by Matas *et al.* [31] as a feature detector that tends to localize blobs. An *extremal region* has all its intensity values greater (or less) than the outer region boundary pixels. Then, a sequence of nested extremal regions is considered. Along this sequence scale change between neighboring regions is estimated. Maximally stable are the local minima of this quantity. In this fashion, nested regions are also likely to appear. The detected regions can have very irregular shapes, therefore an ellipse is fitted to detected regions in order to extract local descriptors. A parameter Δ controls the locality of the scale change computation.

Difference of Gaussians (DoG) was originally used for local feature detection by Lowe [29]. DoG is an efficient way to approximate the Laplacian of Gaussian and to detect edges at various image scales. A Gaussian kernel is used to create multiple blurred versions of the image per octave. Simple subtraction of two consecutive blurred images produces the DoG response. Interest points are located in the scale-space as local-maxima in a 3D search area of size 3.

The original SIFT detection algorithm [29] employs DoG, but further applies additional steps to filter out points belonging to edges or low-contrast regions. However, in our experiments with DoG detector, we keep all the points for a denser representation.

B. Dense patch sampling

Regular grid dense sampling. In contrast to interest points, dense sampling methods give less importance to high repeatability and try to provide a dense coverage of the depicted objects. The most popular method is to sample points on a regular grid, every δ_{xy} pixels. Depending on the application δ_{xy} can be really small, such as 3, or quite larger, such as 16. In order to provide some scale tolerance, different scales are considered by following the same procedure at n_σ multiple scales of the image. All the patches from different scales are pooled together in the end. A typical value for n_σ is 5, with 2 scales per octave. We adopt this choice in our experiments.

Dense interest points were introduced by Tuytelaars [53] as a hybrid solution to trade-off between sparse interest point detection and dense sampling. Instead of selecting the center pixel of grid cell, as in regular dense sampling, they conduct local search inside each cell over spatial and scale space. The point with maximum response is kept per cell. Selected points are not necessarily local maxima. Since a single interest point is selected from each cell, patches are very likely to be localized on smooth regions. Moreover, quite a few patches are localized on the cell borders. In the original work, a large local search area of 16×16 pixels $\times$ 8 scale levels is used, while in our study we experiment with smaller search areas in order to supply more points.

IV. PROPOSED METHODS

Our purpose is to provide a dense set of patches without targeting high repeatability. We rather focus on accuracy in retrieval and fine-grained classification. The proposed methods try to provide a nice coverage of the depicted objects, while focusing on edges in addition to corners and blobs.

In this section we present the proposed approaches for dense patch representation. We relax the Harris-Laplace detector to detect mostly edges. We propose to use convolution by Zernike polynomials [60] as a response function to select point locations. Moreover, we modify the well-known region detectors, such as MSER, to sample points on edge maps, or the edges of the detected regions. Finally, we propose a detector that exploits the ℓ_2-norm of already extracted local descriptors to select the patches. In the following, whenever we mention that local maxima search is performed, a 3×3 neighborhood is used.

A. Relaxing Harris-Laplace detector

Harris-Laplace detector is designed to detect corners. We propose to replace Harris-Laplace cornerness function with Frobenius norm in order to select other points in addition to corners. We also propose to adjust the local extrema selection criterion to sample denser points for classification context.

The first modification is to adjust the standard cornerness function to sample denser points. We estimate the energy of the Harris matrix by its *Frobenius norm*. It is defined as the square root of the sum of the absolute squares of its elements. This offers the ability to produce high response for structures with high energy other than corners.

As our second modification, we propose to modify the local maxima criterion. Regularly, keypoints are selected as local maxima of some interestingness measure. Instead of using the 8-neighborhood of a pixel as local maxima criterion as the standard procedure, we propose to use multiple 2-neighborhoods in different directions, as shown in Figure 4. We select the points that are local maxima in any of those neighborhoods. In this fashion, we retrieve more points along the edges of objects or object parts.

In order to investigate the contribution of different modifications, we end up with four different methods. These are, the regular Harris-Laplace (Harris), Frobenius norm based detector (Frobenius), Harris-Laplace with relaxed local maxima selection (relaxed-Harris) and Frobenius norm based detector with relaxed local maxima selection (relaxed-Frobenius).

In the example of Figure 3, we observe that points are located on edges as well as corners using the Frobenius norm. The relaxed local maxima criterion also provides more points outlining the shapes on the image.
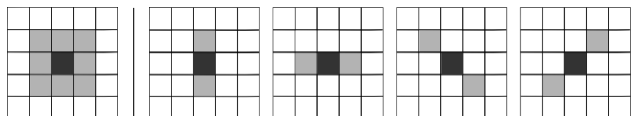

Fig. 4. The search neighborhood typically used to detect local maxima is shown on the left. We relax it by taking the union of extrema obtained from multiple neighborhoods shown on the right.
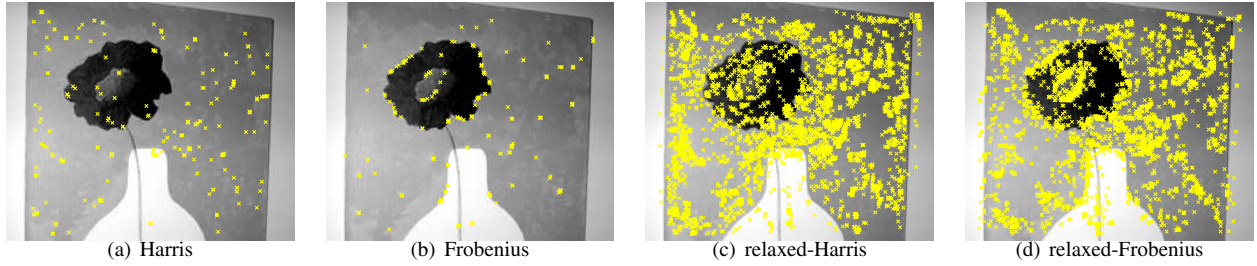
| (a) Harris | (b) Frobenius | (c) relaxed-Harris | (d) relaxed-Frobenius |

Fig. 3. The detected keypoints for our Harris Laplace modifications. We visualize the centers of detected points.

B. Zernike

Zernike polynomials were originally introduced by Zernike [60] and are traditionally used in optics and also as shape descriptors [25][48]. They form a set of orthogonal basis functions defined on the unit disc. We focus on the use of pseudo-Zernike polynomials [7] which have proven to be more robust to image noise [50]. They are defined as

$$v_n^l = R_{nl}(r)e^{il\theta}, \qquad (3)$$

with

$$R_{nl}(r) = \sum_{s=0}^{n-|l|} (-1)^s \frac{(2n+1-s)!}{s!(n-|l|-s)!(n+|l|+1-s)!} r^{n-s}, \qquad (4)$$

where n is the order and l is the repetition. In Figure 5 we present pseudo-Zernike polynomials of order 1 up to order 4. Starting from low frequency patterns they end up being more complex patterns and more localized.

We normalize pseudo-Zernike functions to a 2D rectangular patch (patch width equal to 11 in all our experiments) and convolve the input image using them. Each pseudo-Zernike function is used as a filter and this response constitutes the interestingness measure for detection of point locations. Local maxima and local minima are detected for each filter independently. We now define the maximum number of patches to be extracted per image as N_z. This is used to define the available capacity per filter, in order to retain only the strongest detections. Capacity is uniformly shared among different filters, and also among maxima and minima. We apply the same procedure in $n_\sigma = 5$ scales, similarly to regular sampling. The ratio of capacities of two consecutive scales is 2. That is equal to the ratio of the down-sampled image areas. Filter responses are ranked per filter, and point locations are selected until each filter capacity is filled or until there are no more local extrema left.

We employ N_f filters, which detect complementary structures as being orthogonal and produce high responses for different structures. In contrast to previous feature detectors, we do not focus particularly on detecting corners, blobs or edges and claim that all such structures are useful for image representation. For given capacity, the more filters we use the stronger features we select per filter. In Figure 6, we show the responses produced by various filters. Observe, from the responses and from the filters of Figure 5, that lower order polynomials will detect vertical edges, horizontal edges and
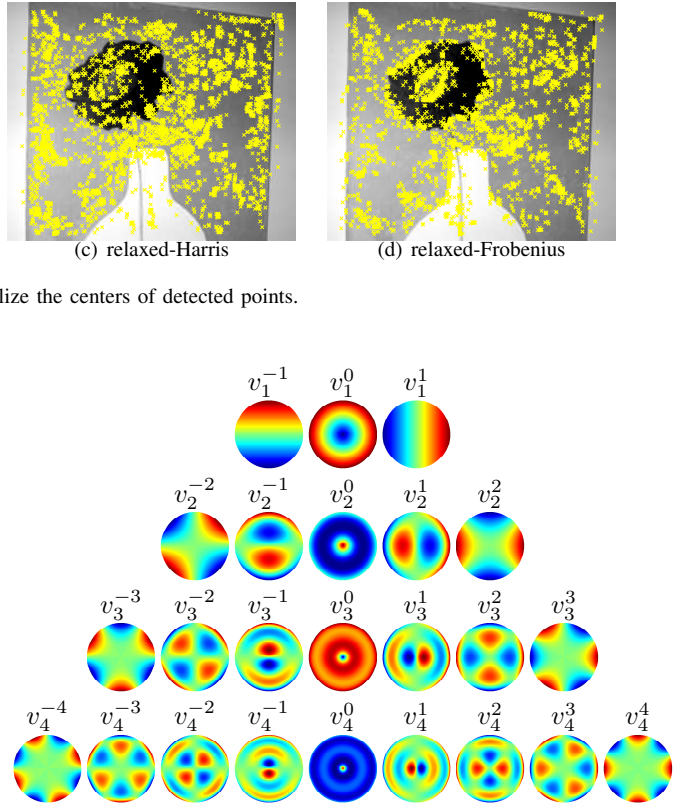

Fig. 5. Visualization of pseudo-Zernike polynomials for $n = 1 \ldots 4$.

blobs (v_1^{-1}, v_1^1 and v_1^0 respectively.), while filters of higher orders will detect more complex image structures.

C. Edge maps

We stress that it is advantageous to describe an image by patches centered on its edges. We do not focus on solving the edge detection problem, where breakthrough research have been done, such as the recent fast edge detection method based on structured forests [13]. We use the existing techniques to sample dense points from images.

We initially consider MSER arbitrarily shaped regions to create an edge map. This 2D map is equal 1 on the region borders and 0 everywhere else. We compute the gradient magnitude on edge pixels and use it as interestingness measure. A local maxima selection is performed to select keypoint locations. Note that only edge pixels are selected in this manner. We center n_σ patches at each detected location, one at each of the multiple scales used. We typically use the same number of scales employed in regular dense sampling, as reported in Section III. We will refer to this detection approach as MSER-edge. Comparing MSER to MSER-edge is a way to compare performance based on region description by its interior or by its boundaries. A similar prior attempt is to describe an MSER region by a set of Local Affine Frames extracted with multiple affine-covariant procedures [39].

We now consider exactly the same procedure, but applied on the edge map derived from the selective search algorithm of Uijlings's *et al.* [54]. Their starting point is Felzenszwalb's segmentation algorithm [16] applied on multiple color spaces.
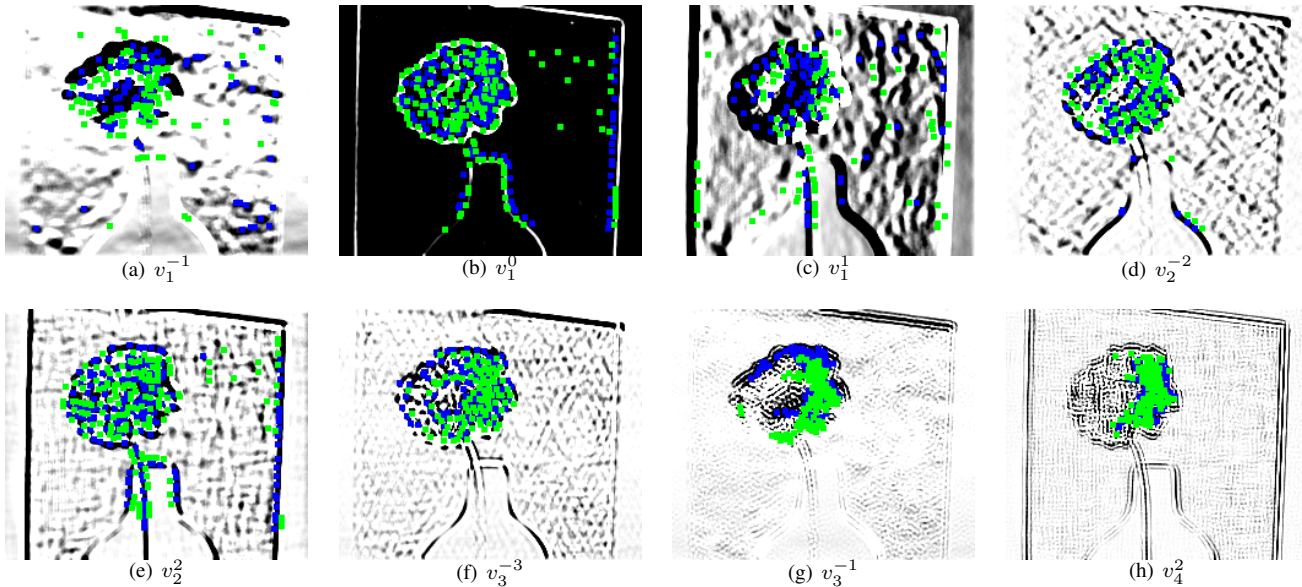
Fig. 6. The convolution response for some of the pseudo-Zernike polynomials. The response map is visualized only for the first scale. We use green (blue) patches to visualize the local maxima (minima). *Best viewed in color.*

Similar segments are merged in a hierarchical manner, and the resulting segments form candidate object regions. We tune it to produce smaller regions, via parameter k, potentially capturing object parts. A super-pixel [1] like segmentation is produced in this fashion.

Except for centering patches on selective search region (SSR) borders, we also fit an upright ellipse to each one. We refer to the latter as SSR, and to the former as SSR-edge.

Finally, we consider the fast edge detection algorithm of Dollár and Zitnick [13] and apply the same process. The only difference is that the interestingness measure now comes directly from the edge detection process, that is the edge strength. We further eliminate all pixels with strength less than τ, in order to control the number of patches per image. We refer to this approach as *fast-edge* in our experiments.

D. Dense ℓ_2-norm local maxima

Gosselin *et al.* [19] show that the classification performance increases when SIFT vectors with low energy are filtered out. This reasoning is supported visually, as these features correspond to homogeneous regions with very little visual information. However, finding a threshold value is not a feasible operation as it requires cross validation. The optimal threshold value may vary depending on the image and the size of sampled patches. The effect of this filtering strategy on regular dense is shown in Figure 7, where the descriptors whose squared ℓ_2-norm is lower than the threshold τ are removed.

It appears that SIFT vectors and in particular their ℓ_2-norm provide an interestingness information subsequent at feature detection. In contrast to traditional detectors that localize interest points by low-level statistics, such as gradient changes, we propose to use the ℓ_2-norm of each SIFT as an interestingness measure. Initially, we compute SIFT on a very dense uniform grid of step $\delta_{xy} = 1$ and at 5 different scales. The ℓ_2-norm

Detector	Section
Dense	Section III-B
Dense Interest Points (Dense-IP) [53]	Section III-B
Dense ℓ_2-norm	Section IV-D
Harris-Laplace [34], +relaxed	Sections III-A, IV-A
Frobenius, +relaxed	Section IV-A
Hessian Affine (HesAff) [35]	Section III-A
Difference of Gaussians (DoG) [29]	Section III-A
Zernike	Section IV-B
MSER [31], +edge	Sections III-A,IV-C
Selective Search [54] (SSR), +edge	Section IV-C
Edge Detector (Fast-edge) [13]	Section IV-C

TABLE I
LIST OF DETECTORS USED IN EXPERIMENTS.

of SIFT descriptors constitutes a response map. That is the interestingness value per pixel. Following the standard detector procedure, we select the local maxima of this response map. Similar to interest point detectors, we introduce a threshold parameter τ, to only retain the stronger points.

As a consequence, patches from homogeneous regions are discarded and the ones with highest energy of SIFT descriptors and more structured regions are retained. As shown in Figure 2 the retained points are mostly localized along edges and corners. It is the only proposed method that is based on the local descriptor to compute the interestingness measure.

V. EXPERIMENTS

We compare different detection strategies in two different applications, image retrieval and fine-grained image classification. In our experiments, we analyze aspects of feature detection and extraction separately and report our findings. List of detectors used in our experiments are shown in Table I. Note that all the detectors we use are rotation variant (up-right).
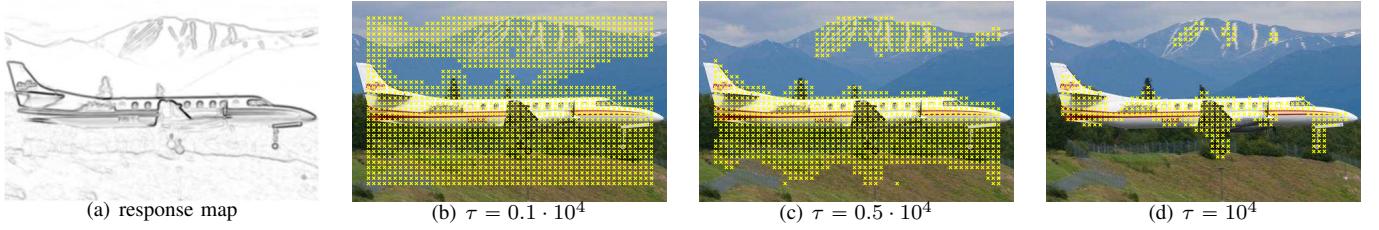
(a) response map   (b) $\tau = 0.1 \cdot 10^4$ (c) $\tau = 0.5 \cdot 10^4$ (d) $\tau = 10^4$

Fig. 7. The response map of ℓ_2-norm of SIFT descriptors is shown in (a). Dark pixels correspond to higher responses. Filtering with different threshold values are shown in (b)-(d). We use the patch size of 41×41 pixels.
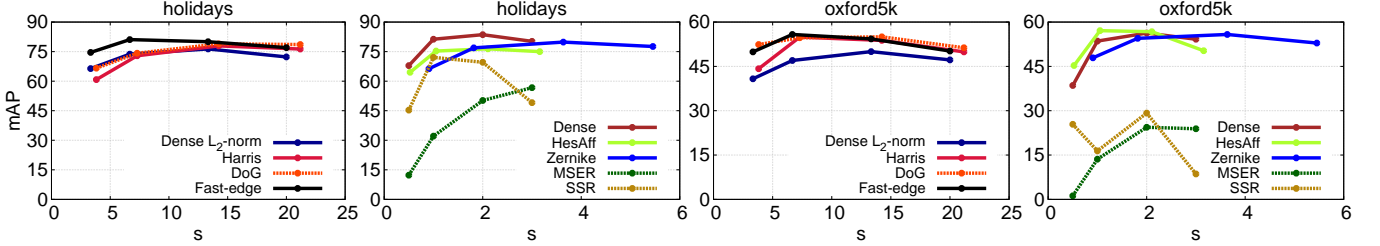

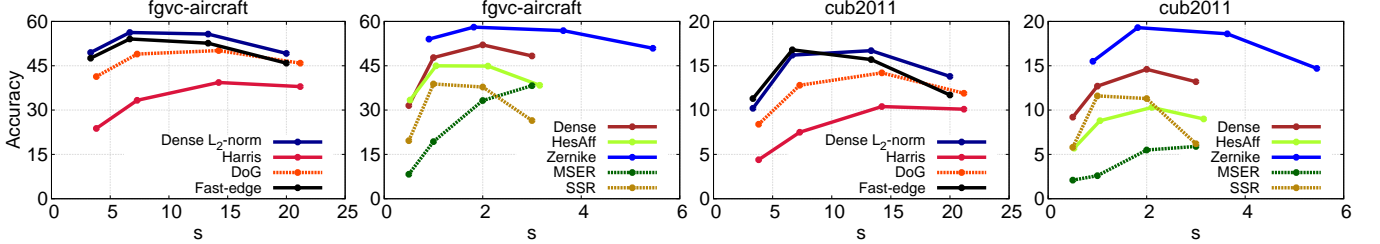Fig. 8. Impact of scaling factor on retrieval performance.

Fig. 9. Impact of scaling factor on fine-grained classification performance.

A. Experimental Setup

Image Retrieval datasets. INRIA Holidays dataset [23] consists of 1,491 personal holiday photographs. There are 500 image groups, with each group having a single query image. We use the up-right version of this dataset, where each image is rotated according to the natural orientation of objects depicted. Flickr60k dataset [23] is used as a training set for this dataset.

Oxford5K Buildings dataset [45] has 5,063 images of Oxford landmarks collected from Flickr. There are 55 query images. Training is performed using Paris 6k dataset [46].

Fine-grained classification datasets. We use the following evaluation benchmarks:

- FGVC-Aircraft dataset [30] has 10,200 aircraft images. There are three different classification tasks, classification by 102 different variants, 70 different families, and 41 different manufacturers. We perform experiments on classification of variants, and use the provided train, validation and test sets.
- Caltech-UCSD Birds 2011 dataset [55] consists of images of bird species. The dataset has 200 categories, and 11,788 images. We randomly select 30% of the training images as validation set [55].
- Oxford-IIIT Pet dataset [40] has 37 categories and 7,349 images of cats and dogs. We randomly select 50% of the training images as validation set [40].

- The 102 Category Flower dataset [37] consists of 102 flower categories, with each category having between 40 and 258 images. We use the provided train, validation and test sets.

We did not make any use of the provided objects regions or segments in any of the aforementioned datasets. Although using segmentation information is shown to improve results [57], we do not consider such an approach. Evaluating the performance on full images offers a direct and fair way to compare different detectors.

Image retrieval pipeline. All images are initially down-sampled to 150k pixels. We extract 128-dimensional SIFT descriptors from the detected set of keypoints. We apply RootSIFT [3], [21], center, rotate along the PCA-axis and ℓ_2 normalize the local descriptors. Codebook of 256 visual words is trained with k-means on an independent dataset. We encode the set of local descriptors with the VLAD [24] representation. Finally, power-law [43] and ℓ_2 normalization is applied to the VLAD vectors. During evaluation, nearest neighbors are found for each query vector, and the performance is measured with mean Average Precision (mAP).

Fine-grained classification pipeline. The down-sampling of images and post-processing of local descriptors is the same as in the retrieval task, except to the fact that we reduce to 80 dimensions using PCA, instead of only rotating and keeping all dimensions. A GMM of 256 components is trained, and the

Detector	conversion	p=11	p=21	p=41	p=61
Dense Dense-IP [53]	$s = (p-1)/20$	0.5	1	2	3
HesAff [35] MSER [31] SSR [54]	$s = (p-1)/20$	0.5	1	2	3
Harris-Laplace [34] DoG [29]	$s = p/2.88$	3.82	7.29	14.24	21.18
Dense ℓ_2-norm MSER-edge SSR-edge Fast-edge [13]	$s = (p-1)/3$	3.33	6.67	13.33	20
Zernike	$s = (p-1)/11$	1	1.91	3.73	5.55

TABLE II
RELATION OF SCALING FACTOR TO PATCH SIZE. WE LINEARLY RELATE
SCALING FACTOR TO PATCHSIZE.

Detector	param	explanation	values
Dense Dense-IP [53]	δ_{xy}	step size	4,8,12,16
Dense ℓ_2-norm Harris-Laplace [34] HesAff [35] DoG [29]	τ	threshold	0,50,100,200 0,50,100,200,300,400 0,50,100,200,300,400 0.1,0.2,0.3
Zernike	N	number of features	1k,5k,10k,16k,20k
MSER [31] MSER-edge	Δ	stable region parameter	3,5,10,15
SSR [54] SSR-edge	k	initial segment size	25,50,100
Fast-edge [13]	τ	edge strength threshold	0,0.1,0.2,0.3

TABLE III
PARAMETERS CONTROLLING NUMBER OF POINTS FOR EACH DETECTOR
AND THEIR CORRESPONDING VALUES.

Fisher vector [42] is adopted to represent an image. Power-law [43] normalization is applied to the Fisher vectors, with its optimal parameter found through cross-validation. A linear classifier is trained using stochastic gradient descent [2], and the accuracy is reported.

B. Impact of the scaling factor

Traditionally, a detected region of interest is mapped to a rectangular patch of width equal to p pixels and a local descriptor is extracted from it. It is beneficial to enlarge this measurement region by a scaling factor [49], denoted by s. The ratio between the descriptor measurement region (used to extract local descriptors) and the detection measurement region (used to compute the interestingness measure to detect keypoints) is exactly the scaling factor s. When $s = 1$, these two are identical.

We investigate the impact of the scaling factor for a variety of detection processes. We linearly relate the scaling factor to the patch size, such that larger regions will be normalized to patches of larger resolution. This exact relation is presented in Table II for all the examined detectors.

Affine regions are isotropically enlarged, and when no enlargement is involved, the patch size is equal to 21. For Zernike, no scaling factor corresponds to descriptor measurement region equal to 11, since this is the filter size we use (see Section IV). For regular dense sampling, where there is no detector measurement region, we have arbitrarily defined the scaling factor to be equal to 1 when the descriptor measurement region size is 21 pixels.

We perform experiments to investigate the impact of the scaling factor and present results in Figures 8 and 9. It is observed that such kind of region enlargement is beneficial for all detectors. Performance increases by increasing the descriptor measurement regions, while for large increase it saturates or even drops. Our conclusions agree with the conclusions of Simonyan *et al.* [49], but we evaluate the scaling factor for a larger number of detectors and both classification and retrieval tasks. We adopt the scaling factor that corresponds to a patch size of 41 for all detectors, as it seems to be a good choice overall. The same stands for detectors not included in the corresponding figures. Note that we do not compare different detectors with each other in this experiment. We investigate the impact of scaling factor for each detector individually.

Each detector produces different number of features per image in this experiment. We consider this a crucial parameter for performance and investigate it in Section V-D.

C. Impact of focusing on edges

In Section IV we proposed detectors that have different localization properties than the existing ones. In particular, we try to focus the local representation on image edges.

In Figure 10, we present the performance measured for Harris-Laplace detector and the relaxed modifications that we proposed. Relaxed-Harris produces more keypoints and consistently improves performance. The relaxation based on Frobenius norm also seems to improve. An exception is the Oxford5K Buildings dataset, where due to particular object matching, original Harris keypoints are localized to have higher repeatability and perform better.

We argue that the typical choice of fitting an ellipse and trying to describe a region by its interior is not necessarily the optimal option for the tasks we consider. Figure 11 shows the results of comparing such a traditional approach to our proposal of extracting local descriptors from patches centered on the region borders. We perform this comparison for MSER regions and selective search regions (SSR). Interestingly, the standard choice does not appear to be the optimal. A larger number of features is obtained when focused on edges, which generally results in better performance. Finally, the region detector based on the fast edge detection algorithm [13] (fast-edge) performs rather well, especially for intermediate number of descriptors per image.

Additionally, we conduct localization experiments for the proposed Zernike detector by directly controlling the number of keypoints via parameter N_z. We also capture more complementary information by increasing the number of filters N_f. We evaluate performance using all polynomials up to order 2, 3 and 4. These correspond to 8, 15 and 24 filters respectively. We present the results in Figure 12. Less filters seem to perform better on retrieval datasets. On the contrary, more filters are needed for fine-grained classification, possibly to capture larger intra class variations that exist. We set N_f equal to 8 for retrieval and 15 for fine-grained classification.
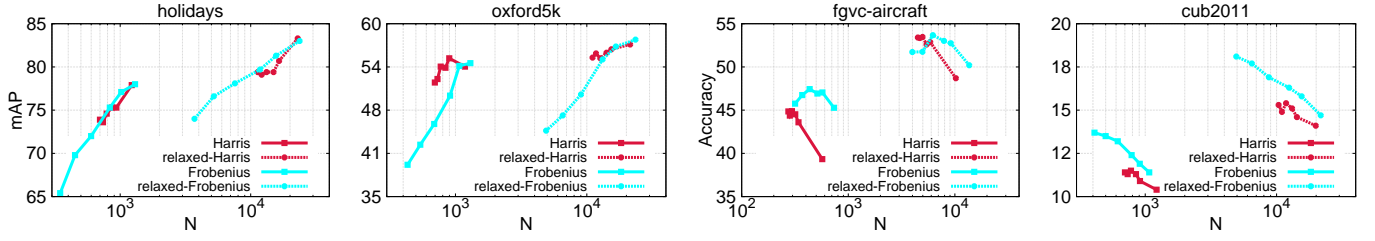
Fig. 10. Performance comparison between Harris-Laplace and its relaxed modifications that we propose. We show performance versus average number of descriptors per image N.
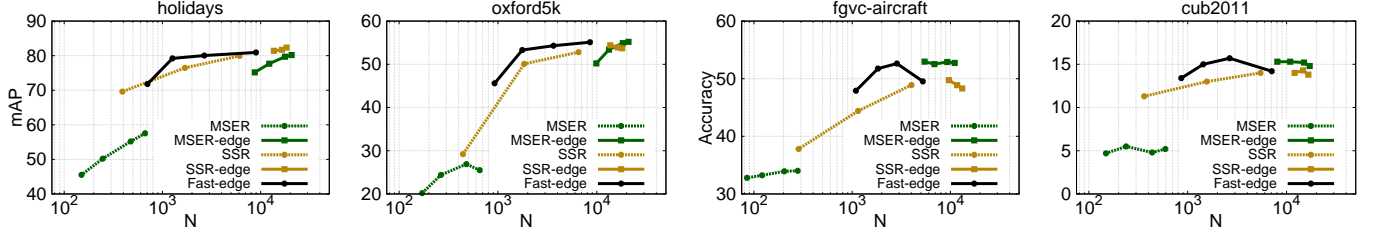

Fig. 11. Performance comparison between approaches that describe a detected region by fitting a single ellipse and approaches that sample patches along the region edges. Fast-edge is another detector that samples patches on edges. We show performance versus average number of descriptors per image N.
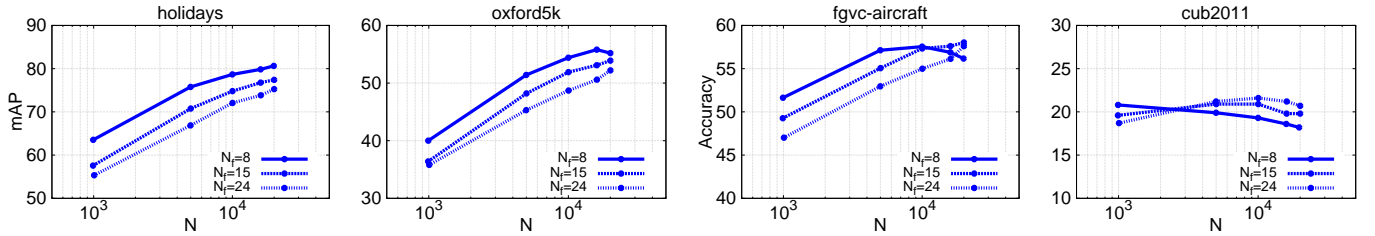

Fig. 12. Performance versus average number of descriptors for Zernike detector with different number of filters.
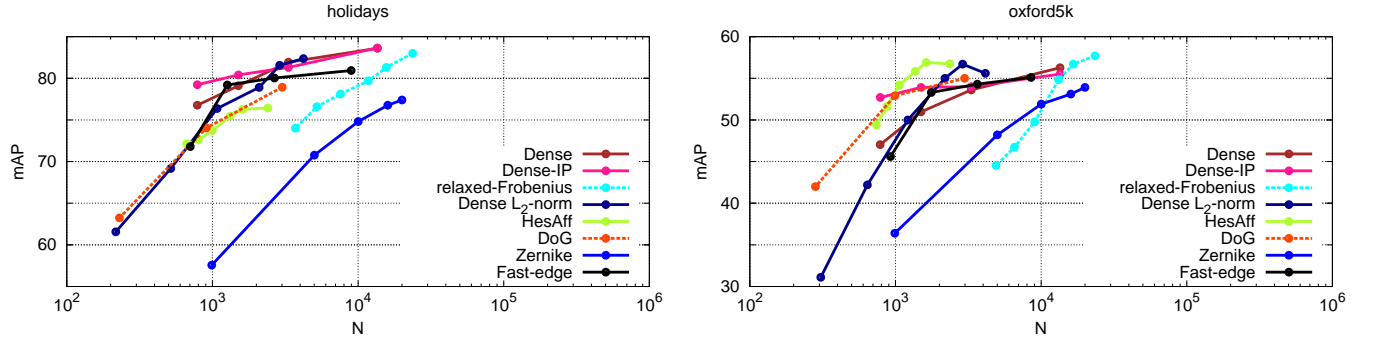

Fig. 13. Performance comparison on image retrieval versus number of descriptors per image N. The parameter values shown in Table III are used.

D. Performance comparison versus number of descriptors

A single parameter for each detector controls the number of keypoints per image. These are shown in Table III. We evaluate performance for multiple values. This allows a fair comparison of all detectors with respect to the number of descriptors used to describe an image. The descriptor set cardinality is directly related to the complexity of encoding stage.

Results of performance comparison between different detectors versus the average number of features per image N are shown in Figures 13 and 14, for retrieval and classification

respectively. Only the best performing variants of the previous experiments are included in this comparison.

On retrieval it appears that the more descriptors the better, while for fine-grained classification this is not always the case. However, such a behavior seems to be consistent among all detectors for a particular dataset. Thus, it seems to be related with the nature of images and objects depicted in them.

The relaxed Frobenius detector achieves the best performance on Oxford5k, however Hessian-Affine also performs very well with less number of features, due to its better
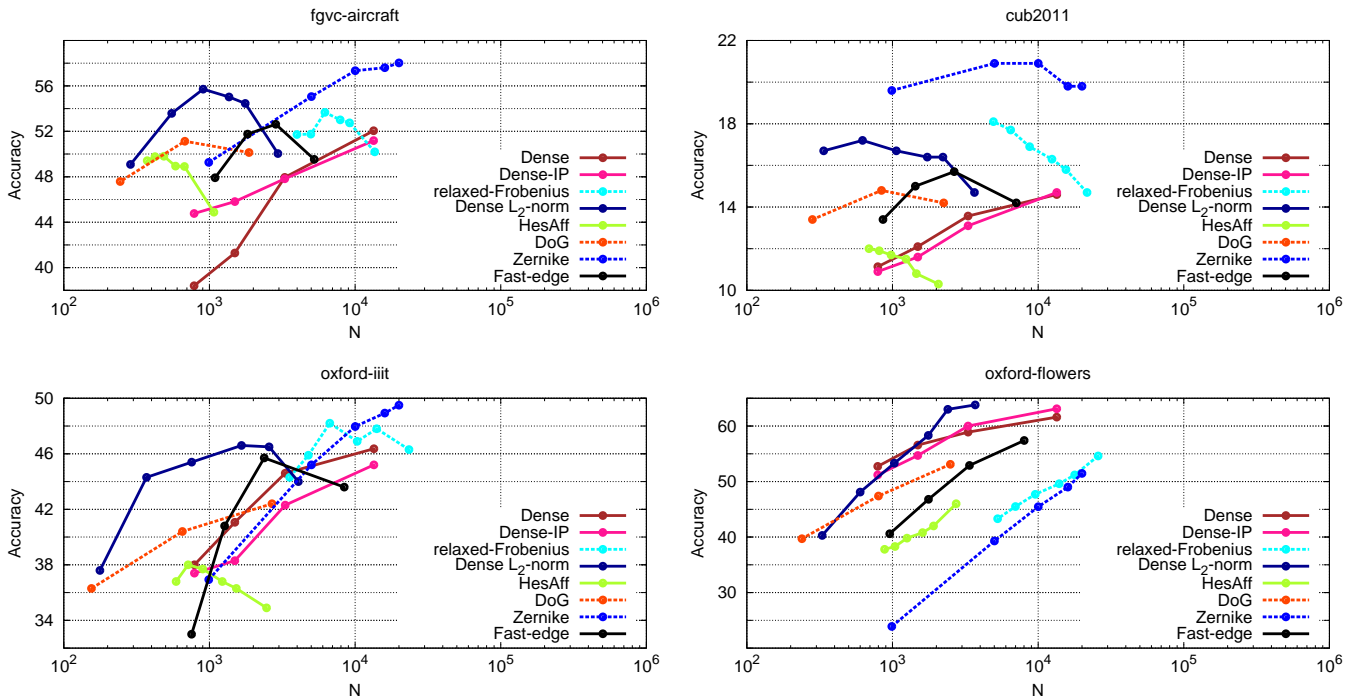
Fig. 14. Performance comparison on fine-grained classification versus number of descriptors per image N. The parameter values shown in Table III are used.

Fig. 15. Detection example showing the failure of Zernike detector on images of the Oxford-Flowers dataset.

Fig. 16. Impact of ℓ_2-norm based filtering when applied on top of a variety of detectors. We show performance versus average number of descriptors per image N. $\ast$ corresponds to detection with ℓ_2-norm filtering.

localization properties. However, we observe that the standard dense and relaxed-Frobenius detectors give the best performance using larger number of features on Holidays. This fact is related to the nature of the dataset; it contains more scenes and not particular objects. On fine-grained classification, the Zernike detector outperforms all other detectors by far, with an exception on Oxford-Flowers dataset. The latter is attributed to many points localized on background texture, as shown in the example of Figure 15. Interestingly, over both tasks our dense ℓ_2-norm local maxima detector achieves best performance for an intermediate amount of points.

Finally, note that detectors which extract descriptors from patches of 5 constant scales perform better than the detectors that detect local maxima in the scale space. It appears that the former attains enough tolerance to scale changes, although it is not scale invariant.
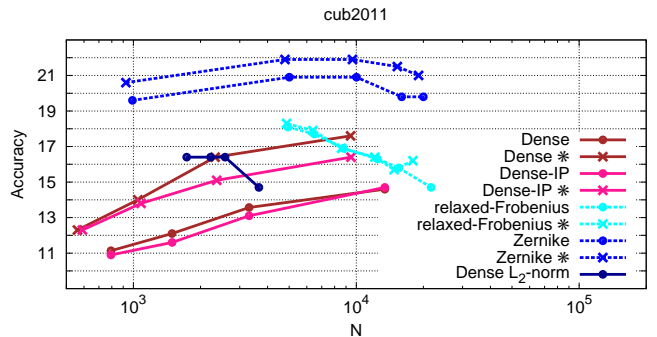
E. Effect of ℓ_2-norm based filtering

Gosselin *et al.* [19] show that the accuracy on fine-grained classification based on regular dense sampling significantly increases when the SIFT vectors with low ℓ_2-norm are filtered out. We apply the same strategy and show results on Figure 16. We compute squared ℓ_2-norm to avoid the calculation of square root, and choose $T = 5000$. As expected, this strategy which aims at removing descriptors extracted from too uniform regions improves performance for regular dense sampling, but the improvement is not significant for detectors that are better localized.

F. State of the art in fine-grained classification

We compare accuracy of Zernike detector, which is best performing among our contributions, with the state of the

Method	#C	FGVC-Air.	CUB2011	Ox.-IIIT	Ox.-Flowers
Zernike	256	56.7	19.3	49.6	51.1
Zernike	4096	66.2	29.6	57.1	51.2
Zernike+SCC	4096	69.9	31.9	59.5	56.1
GMP [36]	256	-	17.0	49.2	73.3
GMP+XColor [36]	256	-	33.3	56.8	84.6

TABLE IV

PERFORMANCE COMPARISON WITH THE STATE OF ART METHODS ON
FINE-GRAINED IMAGE CLASSIFICATION. SCC DENOTES SPATIAL
COORDINATE CODING AND #C IS THE CODEBOOK SIZE. ALL METHODS
USE SIFT DESCRIPTORS, EXCEPT FOR THE LAST ONE WHICH
ADDITIONALLY MAKES USE OF COLOR DESCRIPTORS.

art results on fine-grained classification. Fisher vectors are employed once more to encode an image based on Zernike patches. This time, we also apply horizontal mirroring to images, in order to obtain another Fisher vector representation per image. As recently shown, encoding spatial information can boost performance [19]. We adopt the choice of spatial coordinate coding (SCC) [32][26] that allows the use of much larger codebooks, without increase of dimensionality, in contrast to spatial pyramid matching. We choose SCC instead of spatial pyramid matching following the work of Gosselin [19], which shows that SCC performs better than spatial pyramid matching for fine-grained image classification. It allows the use of much larger codebooks. In particular, we append regularized 2D coordinates of patches to RootSIFT descriptors before quantizing them. Regularization weights are learned via cross validation. We have set $N_z = 10k$ for the Zernike detector.

Several works focus on a single domain of fine-grained classification, *e.g.* birds, and optimize their methods only for that particular domain. Since this is not the case with our study, we compare with methods that are similar. We present our results in Table IV, and compare them with the work of Murray and Perronnin [36]. Using the same codebook, with no additional features, we outperform their results in 2 out of 3 datasets. They also employ color descriptors [11] which further improves the accuracy. This can complement also for Zernike patches. We also increase the codebook size and combine with SCC that drastically boosts performance.

To our knowledge, there are not any published results for the FGVC-Aircraft dataset yet, except for the initial paper that achieved accuracy equal to 48.69% [30]. For the other datasets, it is possible to improve the results using the provided annotations of bounding boxes or object parts. We do not consider such an option. However, the best reported result on CUB2011 is 85.40% [8], which trains convolutional neural networks (CNN) using such annotations. On the other hand, we are superior to methods which use segmentation information on Oxford-IIIT dataset, such as Wang *et al.* [57] that report 59.29%. In the case of Oxford-Flowers dataset, the highest reported score is 86.6% [47], where the authors also use CNNs without the provided segmentation information.

VI. CONCLUSIONS

We investigated dense keypoint detection solutions that lie in between sparse interest points and dense sampling on a uniform grid. We propose to modify existing interest point detectors by relaxing the cornerness criterion and the local maxima selection. We introduce a new detection method using Zernike filters, which provides dense, yet localized image patches. We also show that sampling patches on the borders of a region of interest performs better than the standard choice of fitting an ellipse and describing it by a single descriptor. Finally, we propose to detect dense patches by using the ℓ_2-norm of the descriptor instead of low-level pixel information. To our knowledge, this is the first detection strategy which focuses on descriptors, and the results seem very promising.

Interestingly, solutions employing patches of multiple fixed scales perform better than patches detected as local maxima in the scale space. Albeit not scale invariant, apparently this option provides enough scale tolerance for the tasks of image retrieval and fine-grained classification.

Compared with the existing studies, Zernike patches encoded with a standard technique, such as Fisher vectors, appear to outperform state of the art approaches for some of the fine-grained classification datasets. An exception is the Oxford-Flowers dataset that Zernike seem to perform poorly.

REFERENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *Trans. PAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Good practice in large-scale learning for image classification," *Trans. PAMI*, Jun. 2013.

[3] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *CVPR*, Jun. 2012.

[4] Y. Avrithis and K. Rapantzikos, "The medial feature detector: Stable regions from image boundaries," in *ICCV*, Nov. 2011.

[5] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *ECCV*, 2014.

[6] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, May 2008.

[7] A. Bhatia and E. Wolf, "On the circle polynomials of Zernike and related orthogonal sets," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 50, no. 01. Cambridge Univ Press, 1954, pp. 40–48.

[8] S. Branson, G. V. Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," Arxiv, Tech. Rep., 2014.

[9] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *ECCV*, Oct. 2010.

[10] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod, "CHoG: compressed histogram of gradients," in *CVPR*, Jun. 2009.

[11] S. Clinchant, G. Csurka, F. Perronnin, and J.-M. Renders, "XRCEs participation to imageval," in *ImageEval workshop at CVIR*, 2007.

[12] J. Delhumeau, P.-H. Gosselin, H. Jégou, and P. Pérez, "Revisiting the VLAD image representation," in *ACM Multimedia*, Oct. 2013.

[13] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *ICCV*. IEEE, Dec. 2013, pp. 1841–1848.

[14] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *ICML*, 2014.

[15] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *CVPR*, 2005.

[16] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, vol. 59, no. 2, pp. 167–181, 2004.

[17] M. A. Fischler and R. C. Bolles, "Random sample consensus," *Communications of ACM*, vol. 6, no. 24, pp. 381–395, 1981.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.

[19] P. Gosselin, N. Murray, H. Jégou, and F. Perronnin, "Revisiting the Fisher vector for fine-grained classification," *Pattern Recognition Letters*, 2014.

[20] C. Harris and M. Stephens, "A combined corner and edge detector." in *Alvey vision conference*, 1988.

[21] M. Jain, R. Benmokhtar, P. Gros, and H. Jégou, "Hamming embedding similarity-based image classification," in *ICMR*, Jun. 2012.

[22] M. Jain, H. Jégou, and P. Bouthemy, "Better motion for better action recognition," in *CVPR*, Jun. 2013.

[23] H. Jégou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *ECCV*, Oct. 2008.

[24] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR*, Jun. 2010.

[25] A. Khotanzad and Y. H. Hong, "Invariant image recognition by Zernike moments," *Trans. PAMI*, vol. 12, no. 5, pp. 489–497, 1990.

[26] P. Koniusz, F. Yan, and K. Mikolajczyk, "Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 479–492, 2013.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[28] T. Lindeberg, "Feature detection with automatic scale selection," *IJCV*, vol. 30, no. 2, pp. 77–116, 1998.

[29] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[30] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," Arxiv, Tech. Rep., 2013.

[31] J. Matas, O. Chum, U. Martin, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *BMVC*, Sep. 2002, pp. 384–393.

[32] S. McCann and D. G. Lowe, "Spatially local coding for object recognition," in *Computer Vision–ACCV 2012*. Springer, 2013, pp. 204–217.

[33] K. Mikolajczyk, A. Zisserman, and C. Schmid, "Shape recognition with edge-based features," in *BMVC*, 2003.

[34] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *IJCV*, vol. 60, no. 1, pp. 63–86, Oct. 2004.

[35] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *IJCV*, vol. 65, no. 1/2, pp. 43–72, Nov. 2005.

[36] N. Murray and F. Perronnin, "Generalized max pooling," in *CVPR*, 2014.

[37] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec. 2008.

[38] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *ECCV*, 2006.

[39] Š. Obdržálek and J. Matas, "Object recognition using local affine frames on maximally stable extremal regions," in *Toward Category-Level Object Recognition*. Springer, 2006, pp. 83–104.

[40] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *CVPR*, 2012.

[41] M. Perdoch, O. Chum, and J. Matas, "Efficient representation of local geometry for large scale object retrieval," in *CVPR*, Jun. 2009.

[42] F. Perronnin and C. R. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR*, Jun. 2007.

[43] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *ECCV*, Sep. 2010.

[44] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, Jun. 2007.

[45] ——, "Oxford5k dataset," http://www.robots.ox.ac.uk/~vgg/data, 2007.

[46] ——, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *CVPR*, 2008.

[47] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," Arxiv, Tech. Rep., 2014.

[48] J. Revaud, G. Lavoué, and A. Baskurt, "Improving Zernike moments comparison for optimal similarity and rotation angle retrieval," *Trans. PAMI*, vol. 31, no. 4, pp. 627–636, 2009.

[49] K. Simonyan, A. Vedaldi, and A. Zisserman, "Learning local feature descriptors using convex optimisation," *Trans. PAMI*, 2014.

[50] C.-H. Teh and R. T. Chin, "On image analysis by the methods of moments," *Trans. PAMI*, vol. 10, no. 4, pp. 496–513, 1988.

[51] G. Tolias, Y. Avrithis, and H. Jégou, "To aggregate or not to aggregate: Selective match kernels for image search," in *ICCV*, Dec. 2013.

[52] G. Tolias and H. Jégou, "Visual query expansion with or without geometry: refining local descriptors by feature aggregation," *Pattern Recognition*, Apr. 2014.

[53] T. Tuytelaars, "Dense interest points," in *CVPR*, 2010.

[54] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.

[55] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[56] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin, "Action recognition by dense trajectories," in *CVPR*, Jun. 2011.

[57] Z. Wang, J. Feng, and S. Yan, "Collaborative linear coding for robust image classification," *IJCV*, pp. 1–12, 2014.

[58] S. Winder and M. Brown, "Learning local image descriptors," in *CVPR*, Jun. 2007.

[59] S. Winder, G. Hua, and M. Brown, "Picking the best Daisy," in *CVPR*, Jun. 2009.

[60] v. F. Zernike, "Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasenkontrastmethode," *Physica*, vol. 1, no. 7, pp. 689–704, 1934.

[61] W. Zhao, H. Jégou, and G. Gravier, "Oriented pooling for dense and non-dense rotation-invariant features," in *BMVC*, Sep. 2013.

Ahmet Iscen received his B.S. degree from The State University of New York at Binghamton, and M.S from Bilkent University. He is a PhD student at Inria Rennes and University of Rennes I, working on large-scale image retrieval.
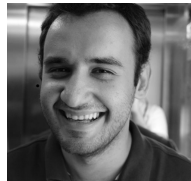
Giorgos Tolias holds a Ph.D. in Electrical and Computer Engineering from the National Technical University of Athens. He is a post-doctoral researcher at Inria Rennes, working mainly in the area of Computer Vision with focus on image search.

Philippe-Henri Gosselin received the PhD degree in image and signal processing in 2005 (Cergy, France). After 2 years of post-doctoral positions at the LIP6 Lab. (Paris, France) and at the ETIS Lab. (Cergy, France), he joined the MIDI Team in the ETIS Lab as an assistant professor, and then promoted to full professor in 2012. His research focuses on machine learning for online multimedia retrieval. He developed several statistical tools for dealing with the special characteristics of content-based multimedia retrieval. This includes studies on kernel functions on histograms, bags and graphs of features, but also weakly supervised semantic learning methods. He is involved in several international research projects, with applications to image, video and 3D objects databases.

Hervé Jégou holds a Ph.D in Computer Science from the University of Rennes, defended in 2005 on joint source channel coding. He joined INRIA as a permanent researcher in 2006, in the LEAR team, and moved to INRIA Rennes in 2009. His research interests include large scale indexing, essentially for large image and video collections.