J. Vis. Commun. Image R. 33 (2015) 301–308

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

SuperPixel based mid-level image description for image recognition ☆

H. Emrah Tasli, Ronan Sicre, Theo Gevers *

Intelligent Systems Lab Amsterdam – Informatics Institute, University of Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 31 March 2015
Accepted 30 September 2015
Available online 26 October 2015

Keywords:
Computer vision
Pattern recognition
Image classification
Image retrieval
Feature extraction
Mid-level cues
Superpixels
Feature encoding

ABSTRACT

This study proposes a mid-level feature descriptor and aims to validate improvement on image classification and retrieval tasks. In this paper, we propose a method to explore the conventional feature extraction techniques in the image classification pipeline from a different perspective where mid-level information is also incorporated in order to obtain a superior scene description. We hypothesize that the commonly used pixel based low-level descriptions are useful but can be improved with the introduction of mid-level region information. Hence, we investigate superpixel based image representation to acquire such mid-level information in order to improve the accuracy. Experimental evaluations on image classification and retrieval tasks are performed in order to validate the proposed hypothesis. We have observed a consistent performance increase in terms of Mean Average Precision (MAP) score for different experimental scenarios and image categories.

1. Introduction

Object recognition is usually defined as the ability to assign labels to objects at multiple conceptual levels, from specific identification to coarse categorization. Possible identity preserving transformations like scaling, rotation, occlusion, changes in intensity, size and pose might be present during the assignment procedure. Ideally, a classification system should provide accurate performance in the presence of such transformations.

Recognizing and localizing semantic objects in a complex scene is a challenging problem that is solved efficiently and successfully by the human visual and cognitive system. However, no method has offered a human-like performance yet. This leads to the following natural question: Where is the "gap" in the image understanding pipeline?

High resolution cameras with good performance under low light conditions and HDR functionality are available. With such equipment, more detailed shots of a scene can be captured compared to bare human eyes; yet, current methods are far beyond the capacity of a basic judgment of a human.

Previous work investigates the perceptual gap between the low-level visual input and the high-level conceptual identification [1]. Studies in neuroscience imply the importance of the feature extraction step for a more accurate visual understanding [2]. The human cognition process is composed of the combinations of complex features [3]. From the computer vision perspective, in an attempt to address these findings, biologically inspired feature descriptions are studied. These approaches aim at exploring possible improvements in the feature extraction step of the image understanding pipeline [4,5].

In this paper, extending [6], the aim is to explore the feature description process by utilizing hierarchical spatial information from mid-level cues in addition to the commonly used pixel descriptors. Therefore, we investigate a superpixel based region descriptor and apply it on object recognition tasks. The region adaptation power of superpixels on the image boundaries, as shown in Fig. 1, make them an ideal candidate for our purposes.

Pixel based descriptors are widely used in object recognition tasks due to their accepted performance for image description [7]. However, the use of middle and higher level descriptors is important for a superior scene characterization. In the proposed method, the aim is to extend the performance of low level descriptors by utilizing middle level region descriptors. The advantage of the proposed adaptation is that it does not require a fixed region size or shape to define the support area of the descriptor. Region shape is adaptive depending on the spatial image characteristics as shown in Fig. 1. The proposed descriptor is based on the superpixel mean color and variance information in the angular spatial neighborhood. Different region and superpixel sizes as shown in Fig. 1 are used to explore possible contributions by fusing spatially different levels of information.

The main contribution of this paper is to propose a novel superpixel based mid-level region descriptor, which can be used in image classification and retrieval tasks. The proposed descriptor

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author.
E-mail address: theo.gevers@uva.nl (T. Gevers).

http://dx.doi.org/10.1016/j.jvcir.2015.09.021
1047-3203/

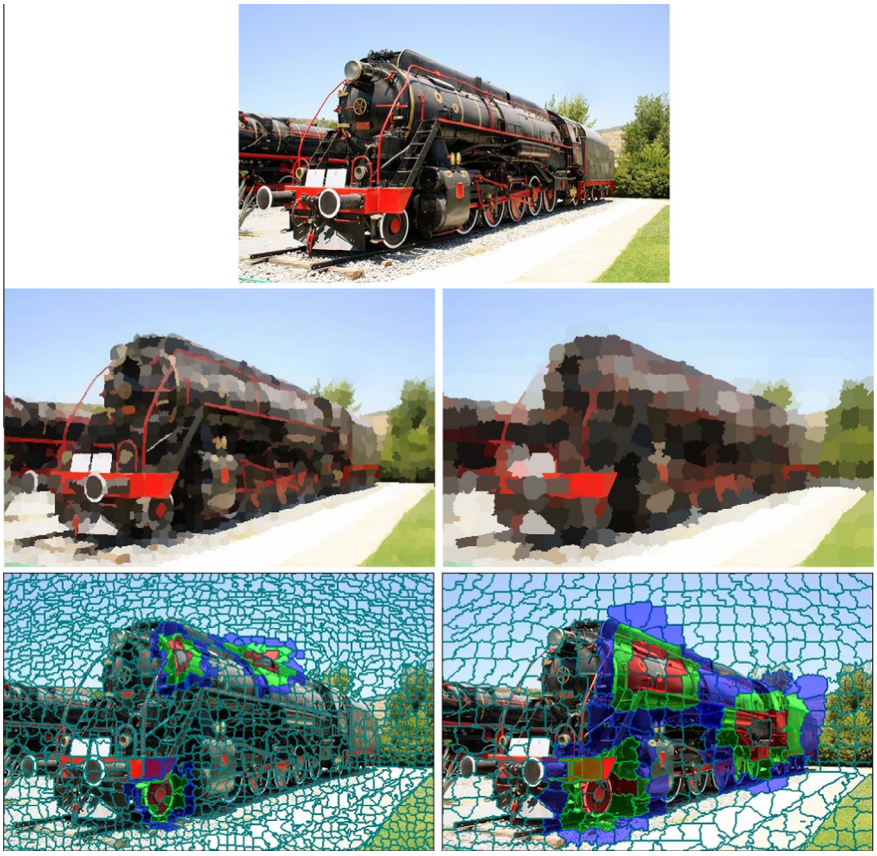

Fig. 1. Describing an image with superpixels. Left: SPs with size 10 × 10 SP. Right: 20 × 20 SP. From top to bottom: Original image; Mean RGB values for each SP region; first (red), second (green) and third (blue) order neighborhoods of randomly selected 3 SP regions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

encodes the intermediate information that is spatially adaptive and hierarchical. Moreover, the proposed descriptor accumulates complementary information w.r.t. the local pixel-level based descriptors, such as SIFT as experimented in this study.

The rest of the paper is organized as follows. Related work and motivation are presented in Section 2. Section 3 provides details on the construction of the superpixel descriptors. The region adaptation idea using the superpixel patches is also presented in the same section. The image classification and retrieval pipeline and different ways of incorporating the proposed region descriptors and region segments are presented in Section 4. The results are discussed in Section 5 before concluding the paper with final remarks and future directions.

2. Related work and motivation

2.1. Image Classification

Object recognition tasks have been vastly studied in the literature [8–10]. A typical object recognition pipeline consists of four major steps: (1) extraction of local image features, (2) encoding of local image descriptors, (3) pooling of encoded descriptors into a global image descriptor, (4) training and classification of pooled image descriptors for the purpose of object recognition. This paper focuses on exploring the first step where local image features are extracted.

Several studies evaluate the performance of the first step in the pipeline; pixel based shape, color, and texture descriptors [11]. Biological insight is also considered to obtain invariance under various viewing conditions [12]. Other studies propose combining different levels (low–mid–high) of information [13]. The second step of the object recognition pipeline has also been widely addressed. For encoding a set of local descriptors into a single high dimensional feature vector, the Fisher Vector method in [14], achieves state-of-the-art performance. The (third) pooling step is also shown to provide improvements. Especially spatial and feature space pooling techniques have been widely investigated [15,16,9,17]. Concerning the final step of the pipeline, discriminative classifiers like SVM are widely accepted as efficient and accurate in terms of classification performance. Judging from the final performance of the-state-of-the-art [10], there is room for improvement in the pipeline.

2.2. Biological insight

The goal of the studies regarding the semantic gap in the image understanding procedure is to determine where the machines lack accuracy compared to humans. In order to address this issue, the way the brain solves visual object recognition task has been investigated. The fact that half of the primate neocortex is engaged during the visual processing, shows the complexity of the whole recognition process [18]. Moreover, recent studies propose strong evidence that a cascade of computations are engaged in the visual object recognition process [19]. However, the underlying algorithm that produces the final result stays mostly undiscovered.

The focus of this paper is not to investigate the neural implications of visual understanding. However, it is important to emphasize the results of recent studies. These results can be valuable to better understand the object recognition process. The neocortex patterns are known to be activated by at least moderately complex

combinations of visual features [3] and it has been observed that output of the neocortex patters can be very informative for achieving robust and real time visual object categorization [20,3]. The goal of this paper is to develop and analyze extensions in feature description and encoding schemes with exploration of hierarchically classified pixel (low level), region (mid level) and scene based (high level) feature descriptors. With the proposed multi layered and region adaptive approach, we hypothesize that a better information accumulation is possible.

2.3. Mid-level features

Several authors have shown the importance of adding an intermediate representation [21], often referred as the mid-level features or mid-level cues for leveraging the performance. We observe three mains trends on mid-level description in the recent literature: *hand-crafted, learned, and unsupervised features*. A large variety of *learned mid-level features* have been proposed. One of the first method was the Deformable Part Model, proposed by [22]. Improvement have been further achieved by using appearance based clustering and sub-categories [23] and by enforcing steerability and separability of the features [24]. Similarly, semantic attributes [25,26] have received a lot of interest. Within the learned mid-level features techniques, we observe a large variety in the nature of the learning data. While some feature are based on extra training data such as labeled fragments [27], sketch tokens [28] or pre-trained object detectors [29], most methods use a standard split of training and testing data to learn the distinctive features, as the *structural element patch model* [30], the *blocks that shout* [31], or the *discriminative parts* [32]. Moreover, regarding *unsupervised mid-level features*, the work of [33] aims at detecting distinctive patches in an image dataset without any label information. On the other hand, *Hand crafted mid-level features* aim at encapsulating information on groups of pixel such as superpixels [34], patches [35] or segments [36]. These descriptors are computed similarly for any given image and do not require any learning, which is a great advantage for efficient image classification systems. Furthermore, these descriptors can be easily applied to image retrieval, which do not have a proper training and test split; and most of these descriptors can effectively encoded with recent well performing methods, such as the Fisher Vectors. The study in [37] proposes mid-level features for object recognition and presents a detailed analysis on different levels of pooling strategies. They define macro-feature vectors as jointly encoded small neighborhoods of SIFT descriptors. The neighborhoods are defined by a fixed size of squares that encode multiple SIFT descriptor into one as the macro-feature vector. This method pursues a similar spatial information utilization as proposed in our work. However, they use only fixed sized (multiple) square regions independent of the region properties. Our method on the other hand, aims at combining spatial characteristics of the region and encoding it into a descriptor that has flexible and adaptive coverage depending on the spatial region properties. A recent work [21] that investigates the role of local and global information in image classification also focuses on exploring the performance limitations of current techniques. Another study that aims at labeling image regions depending on the similarity of the SP features in the training set is presented in [38]. In that study, scene-level matching with global image descriptors is followed by SP level matching of mid-level features. The study in [13] addresses the low-, mid-, and high-level cues. Individual classifiers are trained on different levels of descriptors and classification outputs are combined for the final decision. Descriptor level grouping has also been addressed in a more recent study [35] where local histograms from larger neighboring regions have shown to improve classification performance. This method uses a fixed neighborhood definition to aggregate the local histograms; whereas, our method proposes a flexible and more natural region description.

In order to define the mid-level image regions, superpixel primitives are utilized in this paper. Superpixels (SP) are defined as small pixel groups in the image that are individually consistent in terms of color and textural similarity [39]. This grouping provides advantages especially for graph based applications. By representing the image by SPs instead of pixels, the graph size greatly reduces and this is crucial for computational efficiency. SPs provide an efficient representation of the image that possesses the local color and textural structure in the region. This supports the assumption that pixels in the same SP belong to the same object or region. SP extraction has been widely utilized in computer vision applications mainly as a pre-processing step in order to simplify the node structure. For SP extraction, several methods have been proposed with different advantages [40–43]. In our paper we use the method in [44,45] mainly due to its computation efficiency and structural segmentation performance. A previous method for efficient representation of the images has been previously studied in [46]. The epitome of an image is defined as its miniature, condensed version containing the essence of the textural and shape properties. Similarly, superpixels can also be seen as an efficient image representation with reduced resolution and information encapsulation property.

3. Mid-level cues from superpixels

This paper aims to explore the feature description of the image classification pipeline by using hierarchically generated spatial information from mid-level cues. Therefore, we investigate a superpixel based region descriptor and apply it on object recognition tasks. The reason of selecting superpixels is the region adaptation power on the object boundaries.

The proposed improvement in the feature extraction step is the utilization of SuperPixel based Angular Differences (SPAD) method. This technique uses the intensity difference between the superpixels in a neighborhood. The angular intensity differences in the SP neighborhoods are accumulated in order to define the region covered by the irregular shaped superpixels.

3.1. Superpixel extraction

For the purpose of our mid-level descriptor, extracted SP patches should possess several structural properties. Firstly, the extraction method preserves local structure by adapting to the local object and region boundaries. Secondly, undersegmentation of the regions is avoided to yield an expressive image representation. Thirdly, regular region identification is targeted with quasi-uniform SP regions. Uniform localization and compactness are required to form regular grid structure among the graph models with unbiased neighbor relations. Finally, computational complexity should be kept to a minimum. Based on these criteria, the method in [44] is selected for our purposes. In order to generate a scalable descriptor, different sizes of SPs are hierarchically extracted based on the initial grid structure ($3\times3,\ 5\times5,\ 10\times10,\ 20\times20$).

3.2. Superpixel neigborhood structure

Each SP patch p corresponds to a node $v\in V$ of an undirected graph $G=(V,E)$. Each edge $e\in E$ of the graph is assigned a weight depending on the similarity of the nodes that it connects. For each SP, the neighborhood of p is defined as N_p^n where n corresponds to the order of the neighborhood with $n\in\{1,2,3\}$ in our implementation. Note that two SPs are neighbors if they share a boundary.

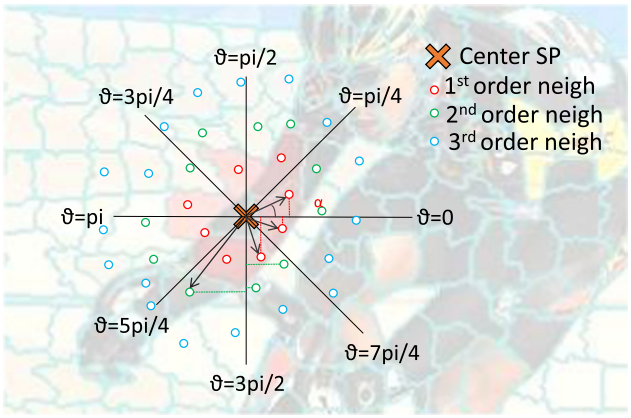

Fig. 2. Computation of angular differences on the superpixel grid. (Best viewed in color) Projection of the closest superpixels are accumulated on the final intensity difference. X represents the central SP and circles in different colors ("red", "green", and "blue") represent the 1st, 2nd, and 3rd order neighbor superpixel centers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The nth order neighborhood is composed of all the neighbors that can only be reached through n hop(s) from a specific SP. For the given parameter settings, we can roughly calculate the region coverage (or surface area) with 3 levels of neighborhood for 20×20 SP size as $(2n+1) \times 20 \rightarrow 140 \times 140$ pixels for $n = 3$. This coverage can be adjusted with different sized SPs or neighborhood levels. In our implementation we use up to the 3rd level of neighborhood with the following SP sizes: $3 \times 3,\ 5 \times 5,\ 10 \times 10,\ 20 \times 20$.

While generating the neighborhood structure, we iterate over all the individual nodes and define the neighborhood relations. The distance metric d_{p,q_i} between the adjacent nodes p and q_i ($q_i \in N_p^n$) is computed as below:

$$d^c_{p,q_i} = e^{\frac{-(\mu^c_p - \mu^c_{q_i})^k sign(\mu^c_p - \mu^c_{q_i})^{k-1}}{\sigma^c_p}},\ k = 1,2, \tag{1}$$

where μ^c is the mean color of the cth index of the color channel. Alternatively, the distance metric $\hat{d}_{p,q_i}$ is evaluated, see Section 4,

$$\hat{d}^c_{p,q_i} = (\mu^c_p - \mu^c_{q_i}) sign(\mu^c_p - \mu^c_{q_i})^k,\ k = 1,2. \tag{2}$$

σ_p is the variance of the mean color values in the nth neighborhood:

$$\sigma^{c2}_p = \frac{1}{\|N^n_p\|} \sum_{i=1:\|N^n_p\|} (\mu^c_p - \mu^c_{q_i})^2, \tag{3}$$

where $\|N^n_p\|$ is the total number of neighbors of the SP p within the nth neighborhood.

In order to compute the angular difference, the angular orientation of each SP with respect to the central SP is required. The angular orientation $arg(p,q_i)$ (argument of the vector $(\vec{p} - \vec{q_i})$ in R^2) between the adjacent nodes p and q_i ($q_i \in N_p^n$) is computed as:

$$arg(p,q_i) = \begin{cases} \arctan(\frac{(p^y - q_i^y)}{(p^x - q_i^x)}) & \text{if } x \geqslant 0 \\ \arctan(\frac{(p^y - q_i^y)}{(p^x - q_i^x)}) + \pi & \text{if } x < 0 y \geqslant 0 \\ \arctan(\frac{(p^y - q_i^y)}{(p^x - q_i^x)}) - \pi & \text{if } x < 0 y < 0 \end{cases} \tag{4}$$

where $p^x,\ p^y$ correspond to the x and y pixel coordinates of the SP p.

3.3. Superpixel based Angular Differences (SPAD)

The generates superpixels and the neighboring relations are used for the proposed mid-level descriptor. Fig. 2 presents the proposed idea where central and neighboring SPs are generated in a realistic configuration for illustration purposes.

The coverage of the neighborhood depends on the size of the extracted SP and the number of neighbor levels. Local SP neighborhood in Figs. 1 and 3 shows the extracted SP boundaries on the original image. On the colored area, the different orders of neighborhoods of the central SP are emphasized with "red", "green" and "blue" colors.

The extracted superpixels and the neighborhood structure are used to compute the angular intensity differences and variances for different (1st, 2nd, and 3rd) levels of neighborhood in Section 3.3. This step is followed by the fusion of the computed angular differences for different sizes of superpixels.

3.3.1. Angular difference computation

We divide the angular space in 8 equal bins to compute the intensity differences of superpixels for different orders of neighborhood. Fig. 2 illustrates the proposed idea where different colored centers contribute to the intensity difference term in the 8 bin angular orientations.

D^c_θ is the angular intensity difference between the center SP p and its neighbors at the selected angle θ and color channel c. In our implementation, we use $N_\Theta = 8$ bin orientations where $\theta \in \Theta = \{0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4\}$. The angular difference D^c_θ is computed as the summation of the projection of all the SPs assigned to this specific bin, as well as the projection of the SPs assigned to his 2 direct neighbors. SPs are assigned to their three closest (in terms of angular orientation) bins, following (5). Fig. 2 shows the projected points for $\theta = 0$ for the 1st neighborhood and $\theta = 3\pi/2$ for the 2nd neighborhood. The dashed lines show the projection of SP centers on the corresponding orientations and intensity differences (positive or negative) are accumulated on each orientation as follows:

$$D^c_\theta = \sum_{i \in Q_\theta} \hat{d}^c_{p,q_i} \cos(arg(p,q_i) - \theta), \tag{5}$$

where the q_i are assigned to the three closest bins Q_θ, as follows:

$$\forall i \in N_p, i \in Q_\theta \iff \forall j \in 1,\ldots,N_\Theta,\ \theta \in \text{argmin}_3(|\Theta_j - arg(p,q_i)|) \tag{6}$$

3.3.2. Incorporating second order statistics

In addition to the angular intensity difference, we also incorporate the angular distribution of second order statistics of the SP patches. As in (5), we compute the angular variances in the SP patches as shown below in (7).

$$V^c_\theta = \sum_{q_i, i=1:3} {\sigma^c_{q_i}}^2 \cos(arg(p,q_i) - \theta), \tag{7}$$

where ${\sigma^c_{q_i}}^2$ is the variance of the cth color channel in SP q_i.

3.3.3. Descriptor fusion

The computation of angular difference D^c_θ and angular variance V^c_θ for 8 orientations produce a 8×1 length vector each. In the proposed method, up to 3 levels of neighborhood information are used to generate a 48×1 sized vector for D^c_θ and V^c_θ together. This vector constitutes the final region descriptor for the given hierarchy as illustrated in Fig. 3 for different orders of neighborhoods and SP sizes.

Different sizes of SPs are used to obtain scale invariance and cover distinct mid-level region cues that we aimed for. The final structure of the descriptor when the angular difference and variance are combined is shown below.

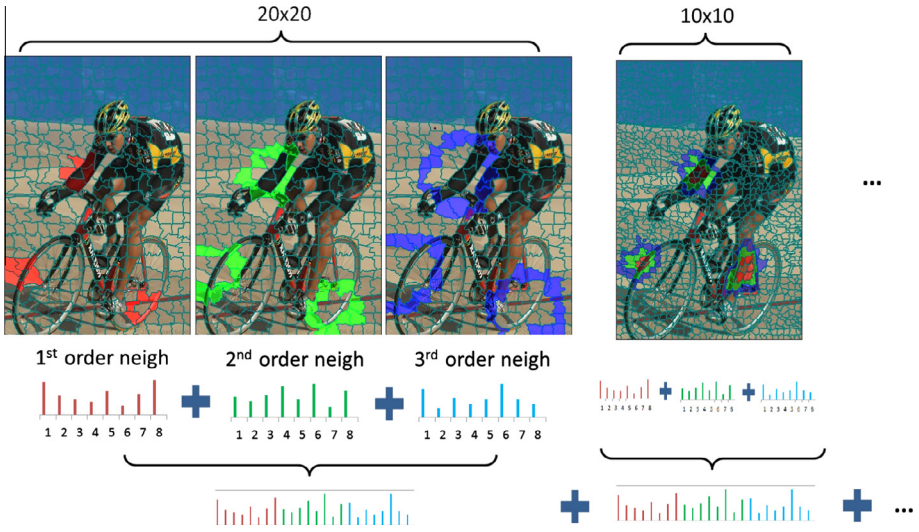

Fig. 3. Angular difference computation. Red, green, and blue colored regions correspond to the 1st, 2nd, and 3rd order neighborhood of the central SP. Angular differences are combined for different neighborhood and SP sizes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$v = [D^n_{\theta_1}\ D^n_{\theta_2}\ \dots\ D^n_{\theta_8}\ V^n_{\theta_1}\ V^n_{\theta_2}\ \dots\ V^n_{\theta_8}]^3_{n=1}$$

As a final step, two descriptors of nth neighborhood D^n_θ and V^n_θ are independently ℓ_2 normalized over all neighborhoods. The normalization step has provided with an increase in the final classification accuracy.

4. Experimental results

The experiments are conducted in a manner to justify the contribution of the proposed mid-level cue combination approach. It is hypothesized that the introduction of mid-level cues at the feature extraction step conceives a complementary information with respect to pixel level information. This has been tested using the image classification pipeline where the proposed superpixel descriptor is combined with the commonly used pixel descriptors. The performance of the proposed approach is further tested on the image retrieval tasks.

4.1. Image classification

In the first part of the experiments, the descriptive performance of SPAD is evaluated on image classification. This task aims at detecting the predefined class of each image in a test set based on training samples. For this purpose, we use the Pascal VOC 2007 Classification Dataset [10], which consist of 9963 images (5011 for training and 4952 for testing). Some examples of the 20 classes in the dataset are: person, motorbike, air plane, cat, cow, bottle, sofa, etc. The measure used to evaluate the performance of a given system is the Average Precision (AP) metric. The Mean Average Precision (MAP) is the averaged AP over all the classes tested.

4.1.1. Classification pipeline

We follow the conventional image classification pipeline presented in [47]. In the first step, mid-level superpixel descriptors are densely extracted from the image. We use the VLFeat toolbox [48] to compute the SIFT descriptors and reduce the dimension of the SIFT features to 64, by using principal component analysis (PCA).

Encoding of the local image descriptors is achieved using the Fisher Vectors (FV). This method is proven to outperform other encoding methods on various tasks such as classification [47]. FVs partition the space using a Gaussian Mixture Model (GMM) and propose the use of first and second order statistics of the difference between the image feature data and the GMM to describe images.

Finally, the training and classification is achieved using linear Support Vector Machine (SVM), as it is shown to perform well with Fisher vector encoding. SVM is trained independently in a one-vs-rest fashion for each image class. Test scores are ranked depending on the output likelihood of each image to belong to the classes in the training set.

We include the proposed method in this pipeline by modifying the feature extraction process. SPADs are computed on each image instead of dense SIFT descriptors. The remaining parts of the pipeline are kept similar; SPADs are encoded with the Fisher Vectors method and SVM is utilized for classification.

4.1.2. Classification results

SPs used in these experiments are extracted based on different grid sizes: 3×3, 5×5, 10×10, and 20×20, see Fig. 3. The SPAD descriptor is computed hierarchically on different scales (SPAD3, SPAD5, SPAD10, SPAD20) for all the images in the dataset.

Construction choices: In this work, various design choices are evaluated. The distance metric chosen: $\hat{d}_p$ (see Eq. (2)) with $k = 2$ is shown to perform better than other possibilities and is more computationally efficient, see Table 1. Next, The number of selected neighborhood in the descriptor is set to $n = 3$. Although $n = 4$ offers a slight improvement, we selected $n = 3$ as it allows a faster computation and smaller descriptor, see Table 2. Finally, we evaluate the performance the distance metric D and the variance V. As we see in Table 3, the variance offers a complementary information to the distance resulting in a large improvement when the two measures are combined.

Fusions of different scales: The MAP scores for each SP are calculated individually as shown in the first four rows in Table 4. The

Table 1
Evaluation of the distance metric d_p and $\hat{d}_p$ for $k = 1,\ 2$, see Eq. (1). SPs of size 25 are utilized and descriptors are further encoded with Fisher Vectors and $k = 256$ Gaussians.

	d_p	$\hat{d}_p$
$k = 1$	0.168	0.173
$k = 2$	0.160	0.185

Table 2
Evaluation of the number of neighborhood utilized to compute the final descriptor.

Neighborhood order	mAP
N12	0.220
N123	0.237
N1234	0.239

last four rows show the improved accuracy with the combined scales. The combination of the descriptors is achieved by *early-fusion* or *mid-fusion* methods. *Early-fusion* encodes all scales of SPAD together and generates a single fisher vector; whereas, *mid-fusion* concatenates the Fisher Vectors of each scales encoded separately. We note that the concatenation step in *mid-fusion* method results in a larger image descriptor compared to *early-fusion*.

Table 4 reveals an improvement in the performance as the SP scale decreases with the increased and finer details. This is expected since the lower scales contain only a rough representation of the image as seen in Fig. 3. This is in accordance with the hypothesis that the lower level pixel information is already well captured with SIFT-like descriptors and we would like to obtain the mid-level additional information that is available in the proposed SPAD descriptors. Moreover, combinations of all scales offer even better results since several levels of region information is incorporated in the combined features. We also observe that *mid-fusion* outperforms *early-fusion* in terms of the MAP score.

Comparison with SIFT: The MAP scores using only the SPAD descriptors are observed to be inferior compared to the state-of-the-art. This is mainly because SP representation is analogous to downscaling the image and running the classification on a lower resolution image. Therefore, in order to obtain a fair comparison between the SPAD and SIFT, we reduced the image dimension so each pixel used to compute SIFT would cover a surface area similar to the area used for SPAD. Therefore, the image is downscaled by 3, 5, and 10 before computing SIFT to obtain the Mini-SIFT (M-SIFT) descriptors. Table 5 shows the individual and combined performance of M-SIFT and SPAD descriptors for scales of 3 and 5 and 10. On the individual performance M-SIFT is observed to be better for the scales 3 and 5, worse of scale 10. However, the combination of these two methods with SIFT, on its original size, shows that SPAD outperforms Mini-Sift for every sizes. This observation validates the initial hypothesis that SPAD could incorporate complementary information so that SIFT benefits largely from SPAD combination.

Finally, we note that the extraction of SPAD is slower than SIFT. SIFT features could be extracted in around 10 h for the Pascal VOC 2007 dataset, on an Intel Xeon 8 cores Desktop and took around 17 h for SPAD. Note that the neighbor search is the longest process and could be optimized. Furthermore, the Fisher encoding and classification is faster due to the smaller dimension of SPAD.

Combination with SIFT: As a final experiment, the proposed SPAD descriptor has been tested against the baseline method where only densely sampled SIFT descriptors are used in the Fisher encoding. Table 6 presents the SIFT baseline MAP score compared with the proposed early and mid fusion of SIFT and SPAD combination.

The increase in the AP scores of the individual classes for the proposed SIFT and SPAD combination is also presented in Fig. 4.

Table 3
Results obtained for mean and variance descriptors for various SPs sizes.

	Mean	Var	Mean & var
SPAD 20	0.194	0.184	0.252
SPAD 10	0.238	0.195	0.300
SPAD 5	0.316	0.233	0.356
SPAD 3	0.353	0.246	0.381

Table 4
SPAD classification MAP scores for Pascal VOC 2007, using Fisher Vectors with $k = 256$ Gaussians. All descriptors are combined using *early-fusion* and *mid-fusion in bold*.

Method	Fisher 256	Dimensions
SPAD 3	0.381	$48 \times 2 \times k$
SPAD 5	0.356	$48 \times 2 \times k$
SPAD 10	0.300	$48 \times 2 \times k$
SPAD 20	0.252	$48 \times 2 \times k$
SPAD 3, 5 Mid	0.406	$2 \times 48 \times 2 \times k$
SPAD 3, 5, 10 Mid	0.417	$3 \times 48 \times 2 \times k$
SPAD 3, 5, 10, 20 Mid	**0.421**	$4 \times 48 \times 2 \times k$
SPAD 3, 5, 10, 20 Early	**0.410**	$48 \times 2 \times k$

Table 5
SPAD and Mini-SIFT classification MAP scores for Pascal VOC 2007, using Fisher Vectors with $k = 256$ Gaussians. Descriptors are further combined with standard dense SIFT using *mid-fusion*.

Method	Fisher 256	Combined with SIFT
SPAD 3	0.381	**0.569**
SPAD 5	0.356	**0.567**
SPAD 10	**0.300**	**0.563**
M-SIFT 3	**0.434**	0.566
M-SIFT 5	**0.378**	0.563
M-SIFT 10	0.281	0.557

Table 6
MAP scores for the standard pipeline and combination with SPAD for Pascal VOC 2007, using Fisher Vectors with $k = \{16, 64, 256\}$ Gaussians. SPAD are combined on the four scales using Early and Mid fusion.

Method	Fisher 16	Fisher 64	Fisher 256
SIFT standard	0.440	0.491	0.549
SIFT & SPAD-Early	0.457	0.514	0.563
SIFT & SPAD-Mid	**0.468**	**0.527**	**0.576**

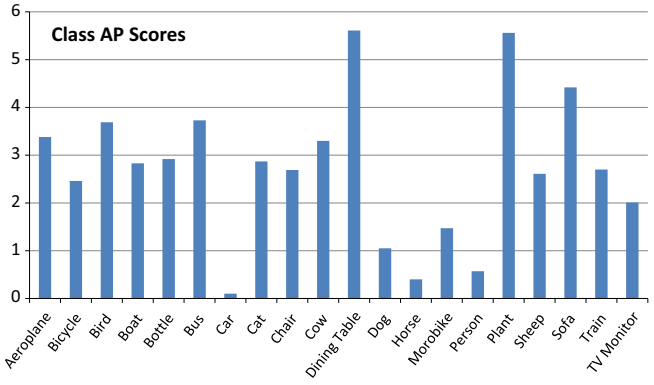

Fig. 4. AP score increase (in percentage) with the proposed SPAD-Mid combination compared to the standard SIFT, for individual classes of Pascal VOC.

The increase in AP score in different classes varies between 0.1% and 5.6%. On the large majority of the classes, we observe a very stable improvement: between 2% and 4%. Increase is obtained regardless of the nature of the data, due to the adaptivity of the proposed descriptor. This observation supports our hypothesis concerning the information gained by utilizing the mid-level cues.

4.2. Image retrieval

In this section, the evaluation of SPAD for the image retrieval task is performed. The aim is to retrieve all samples of a specific query object in an image dataset. The Holidays dataset [49] is used

Table 7
SPAD retrieval MAP scores for Holidays, using Fisher Vectors with $k = 256$ Gaussians.

Method	Fisher 256	Dimensions
SPAD 3	0.587	$48 \times 2 \times k$
SPAD 5	0.592	$48 \times 2 \times k$
SPAD 10	0.581	$48 \times 2 \times k$
SPAD 20	0.552	$48 \times 2 \times k$
SPAD 3, 5, 10, 20 Mid	0.626	$4 \times 48 \times 2 \times k$
SPAD 3, 5, 10, 20 Early	0.614	$48 \times 2 \times k$
SIFT on SPs	0.630	$64 \times 2 \times k$
SIFT & SPAD-Early	**0.663**	$(64 + 48) \times 2 \times k$
SIFT & SPAD-Mid	**0.662**	$(64 + 4 \times 48) \times 2 \times k$
Jegou et al. [50]	0.610	$128 \times k$

in the evaluation. The Holidays dataset consists of 1491 high resolution personal photos of various locations and objects. 500 images are used as query samples in the experiments. The performance is computed similarly by the Mean Average Precision (MAP) score.

4.2.1. Retrieval pipeline

The generic image retrieval pipeline is composed of the following sub-processes: (1) Local image feature extraction. (2) Encoding of the local image descriptors. (3) Image ranking based on the descriptor similarities. In our evaluation, we follow the pipeline proposed by [50]. A dense selection of points for SIFT descriptor extraction performed in the first step. The descriptors are then encoded using Fisher Vectors. Finally, the descriptor distance is computed between the query and the test image from the database using the Euclidean distance of the Fisher vector.

4.2.2. Retrieval results

In terms of the resulting performance, replacement of SIFT descriptors with the proposed SPAD descriptor is evaluated. Furthermore, combination of SIFT descriptors on each superpixels center with SPAD is also tested as shown in Table 7. Finally, the results are compared with a recent work by Jégou et al. [50]. The experimental evaluations show that the MAP scores obtained with *early-fusion* and *mid-fusion* are very similar for the retrieval case. Image description using SIFT is shown to benefit from the proposed SPAD combination, with an increase of 3.2% in MAP score as shown in Table 7.

5. Discussion

This study focuses on the hypothesis that tasks of object recognition and image retrieval can be improved by exploring the limitations at the feature extraction step. Current low-level image descriptors are widely explored for such purposes; however, utilization of mid-level cues can capture additional spatial information. Previous mid-level techniques mostly define a fixed image region and accumulate the low-level information in this predefined window. Different scales of the SIFT descriptor can also collect information from a larger but fixed sized area on the image. On the contrary, we propose a descriptor in the SP domain and define the regions according to the spatial characteristics of the image. The advantage of such an approach is to incorporate region specific information in the descriptor. One can observe the similarities of the proposed method with the LBP descriptor [51], especially in the hierarchical neighborhood idea. However, the two techniques differ in many aspects. The LBP method stores the sign of the differences in the predefined locations of the image. The binary vectors of the sign differences are then accumulated in a histogram on a predefined window. Our method, on the other hand, stores not only the sign but also the magnitude of the difference. Moreover, the shape and size adaptive region coverage makes the proposed method stronger as a spatial region descriptor.

The experimental results show supporting evidence that the proposed method is useful for improving the performance of the object recognition and image retrieval task. The initial hypothesis that the proposed region adaptation could capture additional information is validated with the experiments and could further be extended to improve the performance of any other pipeline where SIFT based feature description is required.

6. Conclusion

This paper focuses on the image recognition task with an emphasis on the feature extraction. We explore the conventional feature extraction techniques from the perspective that mid-level information can be incorporated in this step to obtain a superior scene description. We hypothesize that pixel based low-level descriptions are useful but can be further improved with the introduction of mid-level region information. Thus, we propose a novel descriptor that encapsulates the mid-level information based on SP structure. Image regions are described by computing the oriented mean differences between a central superpixel and its various orders of neighborhood. The variance of the neighbors is further included for a better description. The performance of the proposed descriptor is evaluated on the image classification and retrieval tasks. For the experimental evaluations, baseline score is achieved using SIFT descriptors and we observe 2.7% and 3.2% MAP improvements over the baseline on classification and retrieval tasks, respectively. Based on the experimental evaluations, we could verify our hypothesis that mid-level cues enrich the image description and improve the performance of low-level cues.

References

[1] P.L.G. Martens, R. Walle, Bridging the semantic gap using human vision system inspired features, Self-Organ. Maps (2010) (Chapter 16).
[2] N. Pinto, D.D.J. Doukhan, D. Cox, A high-throughput screening approach to discovering good forms of biologically inspired visual representation, PLoS Comput. Biol. 5 (11) (2009).
[3] N. Rust, J. DiCarlo, Ambiguity and invariance: two fundamental challenges for visual processing, J. Neurosci. 20 (3) (2010) 382–388.
[4] G. Martens, C. Poppe, P. Lambert, R. Walle, Unsupervised texture segmentation using biologically inspired features, in: Workshop on Multimedia Signal Processing, 2008.
[5] J. Zhang, Y. Barhomi, T. Serre, A new biologically inspired color image descriptor, in: ECCV, 2012.
[6] R. Sicre, H. Emrah Tasli, T. Gevers, Superpixel based angular differences as a mid-level image descriptor, in: 22nd International Conference on Pattern Recognition (ICPR), IEEE, 2014, pp. 3732–3737.
[7] D. Lowe, Object recognition from local scale-invariant features, in: International Conference Computer Vision, 1999.
[8] A. Torralba, W. Murphy, K.P. amd Freeman, M.A. Rubin, Context based vision system for place and object recognition, in: Proceedings of International Conference on Computer Vision, 2003.
[9] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Computer Vision and Pattern Recognition, 2006.
[10] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2007 Results, 2007.
[11] N. Pinto, Y. Barhomi, D. Cox, J. DiCarlo, Comparing state-of-the-art visual features on invariant object recognition tasks, in: Workshop on Applications of Computer Vision (WACV), 2011.
[12] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, T. Poggio, Robust object recognition with cortex-like mechanisms, Trans. Pattern Anal. Mach. Intell. 29 (3) (2007) 411–426.
[13] S. Zheng, Z. Tu, A. Yuille, Detecting object boundaries using low-, mid-, and high-level information, in: Computer Vision and Pattern Recognition, 2007.
[14] F. Perronnin, J. Sanchez, T. Mensink, Improving the fisher kernel for large-scale image classification, in: European Conference on Computer Vision, 2010.
[15] Y. Boureau, N. Le Roux, F. Bach, J. Ponce, Y. LeCun, Ask the locals: multi-way local pooling for image recognition, in: ICCV, 2011.
[16] K. Grauman, T. Darrell, The pyramid match kernel: discriminative classification with sets of image features, in: ICCV, 2005.
[17] H.E. Tasli, R. Sicre, T. Gevers, et al., Geometry-constrained spatial pyramid adaptation for image classification, in: IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 1051–1055.
[18] D. Felleman, V. Essen, Distributed hierarchical processing in the primate cerebral cortex, Cereb. Cortex 1 (1) (1991) 1–47.

[19] J. DiCarlo, D. Zoccolan, N. Rust, How does the brain solve visual object recognition?, Neuron 73 (2012)
[20] N. Li, D.D. Cox, D. Zoccolan, J. DiCarlo, What response properties do individual neurons need to underlie position and clutter invariant object recognition, J. Neurophysiol. 102 (1) (2009) 360–376.
[21] D. Parikh, Recognizing jumbled images: the role of local and global information in image classification, in: International Conference on Computer Vision, 2011.
[22] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, Trans. PAMI 32 (2010) 1627–1645.
[23] S.K. Divvala, A.A. Efros, M. Hebert, How important are deformable parts in the deformable parts model?, in: ECCV, Workshops and Demonstrations, Springer, 2012, pp 31–40.
[24] H. Pirsiavash, D. Ramanan, Steerable part models, in: Computer Vision and Pattern Recognition (CVPR), 2012, pp. 3226–3233.
[25] Z. Niu, G. Hua, X. Gao, Q. Tian, Context aware topic model for scene recognition, in: CVPR, 2012, pp. 2743–2750.
[26] Y. Su, F. Jurie, Improving image classification using semantic attributes, Int. J. Comput. Vis. 100 (2012) 59–77.
[27] Z. Liao, A. Farhadi, Y. Wang, I. Endres, D. Forsyth, Building a dictionary of image fragments, in: Computer Vision and Pattern Recognition, 2012, pp. 3442–3449.
[28] J.J. Lim, C.L. Zitnick, P. Dollár, Sketch tokens: a learned mid-level representation for contour and object detection, in: CVPR, 2013.
[29] I. Endres, K.J. Shih, J. Jiaa, D. Hoiem, Learning collections of part models for object recognition, in: CVPR, 2013.
[30] J. Chua, I. Givoni, R. Adams, B. Frey, Learning structural element patch models with hierarchical palettes, in: CVPR, 2012, pp. 2416–2423.
[31] M. Juneja, A. Vedaldi, C.V. Jawahar, A. Zisserman, Blocks that shout: distinctive parts for scene classification, in: CVPR, 2013.
[32] R. Sicre, F. Jurie, Discriminative part model for visual recognition, Comput. Vis. Image Understand. (2015) 1–10.
[33] S. Singh, A. Gupta, A.A. Efros, Unsupervised discovery of mid-level discriminative patches, in: ECCV, 2012, pp. 73–86.
[34] J. Tighe, S. Lazebnik, Superparsing: scalable nonparametric image parsing with superpixels, in: ECCV, 2010.
[35] B. Fernando, E. Fromont, T. Tuytelaars, Effective use of frequent itemset mining for image classification, in: ECCV, 2012.
[36] J. Carreira, R. Caseiro, J. Batista, C. Sminchisescu, Semantic segmentation with second-order pooling, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 430–443.
[37] Y. Boureau, F. Bach, Y. LeCun, J. Ponce, Learning mid-level features for recognition, in: CVPR, 2010.
[38] J. Tighe, S. Lazebnik, Superparsing: scalable nonparametric image parsing with superpixels, in: European Conference on Computer Vision, 2010.
[39] X. Ren, J. Malik, Learning a classification model for segmentation, in: IEEE International Conference on Computer Vision, 2003.
[40] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Susstrunk, Slic superpixels compared to state-of-the-art superpixel methods, PAMI 34 (11) (2012) 2274–2282.
[41] A. Levinshtein, A. Stere, K.N. Kutulakos, D.J. Fleet, S.J. Dickinson, K. Siddiqi, TurboPixels: fast superpixels using geometric flows, IEEE Pattern Anal. Mach. Intell. (2009).
[42] M. van den Bergh, X. Boix, G. Roig, B. de Capitani, L. van Gool, Seeds: superpixels extracted via energy-driven sampling, in: ECCV, 2012.
[43] O. Veksler, Y. Boykov, P. Mehrani, Superpixels and supervoxels in an energy optimization framework, in: Perspectives in Neural Computing, 2010.
[44] H.E. Tasli, C. Cigla, T. Gevers, A. Alatan, Super pixel extraction via convexity induced boundary adaptation, in: IEEE International Conference on Multimedia and Expo (ICME), 2013.
[45] H.E. Tasli, C. Cigla, A.A. Alatan, Convexity constrained efficient superpixel and supervoxel extraction, Signal Process.: Image Commun. (2015) 71–85.
[46] N. Jojic, B.J. Frey, A. Kannan, Epitomic analysis of appearance and shape, in: ICCV, 2003.
[47] K. Chatfield, V. Lempitsky, A. Vedaldi, A. Zisserman, The devil is in the details: an evaluation of recent feature encoding methods, in: BMVC, 2011.
[48] A. Vevaldi, B. Fulkerson, Vlfeat an open and portable library of computer vision algorithms, in: Proceedings of ACM International Conference on Multimedia, 2010.
[49] H. Jégou, M. Douze, C. Schmid, Hamming embedding and weak geometric consistency for large scale image search, in: European Conference on Computer Vision, 2008.
[50] H. Jégou, F. Perronnin, M. Douze, C. Schmid, et al., Aggregating local image descriptors into compact codes, IEEE Trans. PAMI 34 (9) (2012) 1704–1716.
[51] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, Trans. PAMI (2002).