SuperPixel based Angular Differences as a mid-level Image Descriptor

Ronan Sicre, H. Emrah Tasli and Theo Gevers
Intelligent Systems Lab Amsterdam
Informatics Institute, University of Amsterdam

Abstract—This paper focuses on the object recognition task and aims at improving the accuracy with an emphasis on the feature extraction step. Feature extraction is widely used in image classification as an initial step in the pipeline. In this paper, we propose a method to explore the conventional feature extraction techniques from the perspective that mid-level information could be 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 classification accuracy. Detailed experimental evaluations on classification and retrieval tasks are performed in order to validate the proposed hypothesis. A consistent increase is observed in the mean average precision (MAP) score for different experimental scenarios and image categories.

I. 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?

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 a combination 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 improvements in the feature extraction step of the image understanding pipeline [4], [5].

In this paper 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, e.g. SIFT [6]. 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 Figure 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 [6]. 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 Figure 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 Figure 1 are used to explore possible contributions by fusing spatially different levels of information.

The rest of the paper is organized as follows. Related work and motivation are presented in Section II. Section III 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 IV, before concluding the paper with final remarks and future directions.

II. RELATED WORK AND MOTIVATION

A. Image Classification

Object recognition tasks have been vastly studied in the literature [7], [8]. 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 [9]. Biological insight is also considered to obtain invariance under various viewing conditions [10]. Other studies propose combining different levels (low - mid - high) of information [11]. The second step of the object recognition pipeline has
Fig. 1. Describing an image with superpixels. Left: SPs with size $10 \times 10$ SP. Right: $20 \times 20$ 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.

Also been widely addressed. For encoding a set of local descriptors into a single high dimensional feature vector, the Fisher Vector method in [12], 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 [13], [7]. 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 [8], there is room for improvement in the pipeline.

B. Mid-Level Cues

In order to define the mid-level image regions, superpixel primitives (SP) are defined as small pixel groups in the image that are individually consistent in terms of color and textural similarity [14]. 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 [15], [16]. In our paper we use the method in [17] mainly due to its computation efficiency and structural segmentation performance.

A previous method for efficient representation of the images has been previously studied in [18]. 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. The study in [19] 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 [20] 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 labelling image regions depending on the similarity of the SP features in the training set is presented in [21]. In that study, scene-level matching with global image descriptors is followed by SP level matching of mid-level features. The study in [21] 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 [22] 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.

III. 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 local neighborhood. The angular intensity differences in the SP neighborhoods are accumulated in order to define the region covered by the irregular shaped superpixels.

A. Superpixel Extraction

For the purpose of our mid-level descriptor, extracted SP patches should possess several structural properties. First, 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 [17] 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 ($3 \times 3$, $5 \times 5$, $10 \times 10$, $20 \times 20$). Details regarding
the superpixel extraction methodology can be found in our previous work [17].

B. Superpixel Neighborhood 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. For the given parameter settings, we can roughly calculate the region coverage with 3 levels of neighborhood for \( 20 \times 20 \) 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 neighbor levels. In our implementation we use up to the 3\textsuperscript{rd} 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. To obtain a color wise distance \( d_{p,q_i} \) between the adjacent nodes \( p \) and \( q_i (q_i \in N_p^n) \), the distance metric is computed over three color channels:

\[
d^c_{p,q_i} = e^{-\frac{(-\mu_p^c - \mu_{q_i}^c)^2}{\sigma_p^2}}, k = 1, 2, \ldots, \frac{\pi}{\theta} - \frac{1}{\theta}, \quad (1)
\]

where \( \mu^c \) is the mean color of the \( c \)th index of the color channel. The parameter \( k \) is tested with two values for comparison, however no significant difference is observed. \( \sigma_p \) is the variance of the mean color values in the \( n \)th neighbor:

\[
\sigma_p^2 = \frac{1}{||N_p^n||} \sum_{i=1:||N_p^n||} (\mu_p^c - \mu_{q_i}^c)^2, \quad (2)
\]

where \( ||N_p^n|| \) is the total number of neighbors of the SP \( p \) within the \( n \)th 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 \( \text{arg}(p,q_i) \) (argument of the vector \( \langle p - q_i \rangle \)) in \( \mathbb{R}^2 \) between the adjacent nodes \( p \) and \( q_i (q_i \in N_p^n) \) is computed as:

\[
\text{arg}(p,q_i) = \begin{cases} 
\arctan\left(\frac{p^y - q_i^y}{p^x - q_i^x}\right) & \text{if } x \geq 0 \\
\arctan\left(\frac{p^y - q_i^y}{p^x - q_i^x}\right) + \pi & \text{if } x < 0 \ y \geq 0 \\
\arctan\left(\frac{p^y - q_i^y}{p^x - q_i^x}\right) - \pi & \text{if } x < 0 \ y < 0 
\end{cases} \quad (3)
\]

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

The calculated distance and angular orientations are used in the next step to compute the angular intensity differences.

C. Superpixel based Angular Differences (SPAD)

The generated superpixels and the neighboring relations are utilized to generate the proposed mid-level descriptor. Figure 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 Figure 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 (1\textsuperscript{st}, 2\textsuperscript{nd}, and 3\textsuperscript{rd}) levels of neighborhood in Section III-C. This step is followed by the fusion of the computed angular differences for different sizes of superpixels.

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

\( D_{p}^{c} \) is the angular intensity difference between the center SP \( p \) and its neighbors at the selected angle \( \theta \) and color channel \( c \). In our implementation, we use 8 bin orientations where \( \theta \in \{0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4\} \). The angular difference \( D_{p}^{c} \) is computed as the summation of the projection of 3 closest (in terms of angular orientation) SPs in the selected neighborhood order (4). Figure 2 shows the projected points for \( \theta = 0 \) for the 1\textsuperscript{st} neighborhood and \( \theta = 3\pi/2 \) for the 2\textsuperscript{nd} 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:

\[
D_{p}^{c} = \sum_{q_i, i=1:3} d_{p,q_i}^c \cos(\text{arg}(p,q_i) - \theta), \quad (4)
\]

where the index of 3 closest neighbors \( q_i \) is selected iteratively as:

\[
i = \arg\min_{j} (|\theta - \text{arg}(p,q_j)|), \quad j \in N_p . \quad (5)
\]
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.

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 (4), we compute the angular variances in the SP patches as shown below in (6).

\[
\begin{align*}
V_c^\theta &= \sum_{q_i, i=1:3} \sigma_{q_i}^c \cdot \cos(\text{arg}(p, q_i) - \theta) , \\
&= \sum_{q_i, i=1:3} \sigma_{q_i}^c \cdot \cos(\text{arg}(p, q_i) - \theta) , \\
\end{align*}
\]

where \(\sigma_{q_i}^c\) is the variance of the \(c^{th}\) color channel in SP \(q_i\).

Different color spaces and different number of color channels could be utilized in the proposed descriptor. However, in order to make a proper comparison with the gray channel SIFT descriptor and also to keep the descriptor length limited we have used only the gray channel information in the rest of the experiments.

Descriptor Fusion:
The computation of angular difference \(D_\theta^c\) and angular variance \(V_\theta^c\) for 8 orientations produce a \(8 \times 1\) length vector each. In the proposed method, up to 3 levels of neighborhood information are used to generate a \(48 \times 1\) sized vector for \(D_\theta^c\) and \(V_\theta^c\) together. This vector constitutes the final region descriptor for the given hierarchy as illustrated in Figure 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.

\[
v = [D_\theta^1 \ D_\theta^2 \ ... \ D_\theta^n \ V_\theta^1 \ V_\theta^2 \ ... \ V_\theta^n]_{n=1}^3
\]

As a final step, two descriptors of \(n^{th}\) neighborhood \(D_\theta^n\) and \(V_\theta^n\) are independently \(\ell_2\) normalized over all neighborhoods. The normalization step has provided with an increase in the final classification accuracy.

IV. 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.

A. 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 [8], which consist of 9,963 images (5,011 for training and 4,952 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
TABLE I. SPAD CLASSIFICATION MAP SCORES FOR PASCAL VOC 2007, USING FISHER VECTORS WITH K=256 GAUSSIANS. DESCRIPTORS ARE COMBINED USING EARLY-FUSION AND MID-FUSION.

<table>
<thead>
<tr>
<th>Method</th>
<th>Fisher 256</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPAD 3</td>
<td>0.381</td>
<td>48 x 2 x k</td>
</tr>
<tr>
<td>SPAD 5</td>
<td>0.356</td>
<td>48 x 2 x k</td>
</tr>
<tr>
<td>SPAD 10</td>
<td>0.300</td>
<td>48 x 2 x k</td>
</tr>
<tr>
<td>SPAD 20</td>
<td>0.252</td>
<td>48 x 2 x k</td>
</tr>
<tr>
<td>SPAD 3,5 Mid</td>
<td>0.406</td>
<td>2 x 48 x 2 x k</td>
</tr>
<tr>
<td>SPAD 3,5,10 Mid</td>
<td>0.417</td>
<td>3 x 48 x 2 x k</td>
</tr>
<tr>
<td>SPAD 3,5,10,20 Mid</td>
<td>0.421</td>
<td>4 x 48 x 2 x k</td>
</tr>
<tr>
<td>SPAD 3,5,10,20 Early</td>
<td>0.410</td>
<td>48 x 2 x k</td>
</tr>
</tbody>
</table>

Average Precision (AP) metric. The Mean Average Precision (MAP) is the averaged AP over all the classes tested.

Classification pipeline:
We follow the conventional image classification pipeline presented in [23]. In the first step, mid-level superpixel descriptors are densely extracted from the image. We use the VLFeat toolbox [24] 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 [23]. 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.

Classification results:
An evaluation of the proposed system using various scales of SPAD is performed and the results are shown in Table I. SPs used in this experiment are extracted based on different grid sizes: 3 x 3, 5 x 5, 10 x 10, and 20 x 20, see Figure 3. The SPAD descriptor is computed hierarchically on different scales (SPAD3, SPAD5, SPAD10, SPAD20) for all the images in the dataset. The MAP scores for each SP are calculated individually as shown in the first four rows in Table I. The last four rows show the improved performance 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 I 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 Figure 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. Finally, we note that the MAP scores using only the SPAD descriptors are observed to be inferior compared to the state-of-the-art SIFT descriptor, see table II. The reason is that SP representation is analogous to downsampling the image and running the classification on a lower resolution image.

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 II presents the SIFT baseline MAP score compared with the proposed early and mid fusion of SIFT and SPAD combination. These observations support our hypothesis concerning the information gained by utilizing the mid-level cues. The combination of dense SIFT descriptors with SPAD offers better performances and we obtain 2.7% improvement in terms of MAP over the baseline.

B. 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 [25] is used in the evaluation. The Holidays dataset consists of 1,491 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.

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 [26]. A dense selection of points for SIFT descriptor extraction performed in the first step. The descriptors are then encoded using Fisher vectors. Finally,
TABLE III. SPAD RETRIEVAL MAP SCORES FOR HOLIDAYS, USING FISHER VECTORS WITH K=256 GAUSSIANS

<table>
<thead>
<tr>
<th>Method</th>
<th>Fisher 256</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPAD 3</td>
<td>0.587</td>
<td>$48 \times 2 \times k$</td>
</tr>
<tr>
<td>SPAD 5</td>
<td>0.592</td>
<td>$48 \times 2 \times k$</td>
</tr>
<tr>
<td>SPAD 10</td>
<td>0.581</td>
<td>$48 \times 2 \times k$</td>
</tr>
<tr>
<td>SPAD 20</td>
<td>0.552</td>
<td>$48 \times 2 \times k$</td>
</tr>
<tr>
<td>SPAD 3,5,10,20 Mid</td>
<td>0.626</td>
<td>$4 \times 48 \times 2 \times k$</td>
</tr>
<tr>
<td>SPAD 3,5,10,20 Early</td>
<td>0.614</td>
<td>$48 \times 2 \times k$</td>
</tr>
<tr>
<td>SIFT on SpS</td>
<td>0.630</td>
<td>$64 \times 2 \times k$</td>
</tr>
<tr>
<td>SIFT &amp; SPAD-Early</td>
<td>0.663</td>
<td>$(64 + 48) \times 2 \times k$</td>
</tr>
<tr>
<td>SIFT &amp; SPAD-Mid</td>
<td>0.662</td>
<td>$(64 + 4 \times 48) \times 2 \times k$</td>
</tr>
</tbody>
</table>

Jegou et al. [26] 0.610 $128 \times k$

the descriptor distance is computed between the query and the test image from the database using the Euclidean distance of the Fisher vector.

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 III. Finally, the results are compared with a recent work by Jegou et al [26]. 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 III.

V. 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