ABSTRACT
This paper proposes a geometry-constrained spatial pyramid adaptation approach for the image classification task. Scene geometry is used as an input parameter for generating the spatial pyramid definitions. The resulting region adaptation is performed in accordance with the predefined geometric guidelines and underlying image characteristics. Using an approximate global geometric correspondence, exploits the idea that images of the same category share a spatial similarity. This assumption is evaluated and justified in an object classification framework, in which generated region segments are used as an enhancement to the widely utilized “spatial pyramid” method. Fixed region pyramids are replaced by the proposed locally coherent geometrically consistent region segments. Performance of the proposed method on object classification framework is evaluated on the 20 class Pascal VOC 2007 dataset. The proposed method shows consistent increase in the mean average precision (MAP) score for different experimental scenarios.

Index Terms— Image classification, spatial pooling, region segmentation

1. INTRODUCTION
Image segmentation and classification have been widely studied in the literature and significant improvements have been obtained. In this paper, we intend to utilize a regional, geometry-imposed, non-semantic image segmentation method in the classification pipeline. We hypothesize that the proposed region definitions could better encapsulate local information compared to the conventional spatial pyramid method and hence improve the classification accuracy.

In the context of image segmentation, no method can yet accurately segment unconstrained generic scenes with human-like performance. Due to the natural difficulty of the problem, researchers have also investigated ways to introduce some prior information in the process. For this purpose human assistance for image segmentation is utilized in which important and valuable information from the user helps to obtain improved object/region segments in a semantically meaningful context. Seminal work of Boykov et al. introduced the methodology to utilize this information in the image segmentation framework [1]. This method is further improved in [2] to obtain more accurate results. Such a prior information for image segmentation is very valuable but rather expensive in terms of required user assistance.

In this paper, we aim to exploit region segmentation in order to obtain natural image layout for the purpose of image classification. In order to do that, we propose a method that segments image regions automatically for a given input geometry. Region clustering assigns pixels in a one-to-one correspondence to the connected image regions. The use of geometry for image classification is inspired by the "spatial pyramid" idea of Lazebnik et al. [3]. In that paper, hard segment boundaries are utilized for creating hierarchical rectangular windows as region segments. The motivation behind such partitioning is to utilize the locality with combination of the global information in frames, ie. an image on the highway is more likely to contain a car than a toaster [3]. With such geometric partitions, one can better utilize the statistical information related to the location of individual segments, ie. “Sky” is usually in the top part of the image.

A previous work on generic pooling for image classification proposes utilization of similarities between image categories [4]. This has been shown to improve the classification scores with the introduced correspondence between the equivalence classes. In this study we emphasize the idea that geometric properties of the regions have a statistical
relevance to image categories. Hence, we propose to better exploit this relation compared to the conventional "Spatial Pyramid" method where hard region assignment is performed on the image.

2. PROPOSED METHOD

Automatic partitioning of images into semantic regions by using the top-down knowledge with bottom-up grouping approach has been widely discussed in the previous literature [5], [6]. User assisted methods are also proposed for such purposes [7], [8], where they all aim to generate semantic segments that correspond to objects or regions. This idea has been shown to improve the classification scores if performed accurately. In contrast to this semantic partitioning idea, we intend to generate coarse region segments. This is performed according to a predefined geometric prior where a convexity constrained energy formulation is used to preserve the shape of the initial geometry. This would yield adaptation of the region boundary while keeping the region geometry intact.

The proposed method initially segments the image into a large number of (~ 600) color wise similar regions. These so called "Superpixels" (SP) of the image are used as the atomic structures for image representation. The final goal is to partition the image into small number of (3 or 4) spatially coherent regions by dynamically moving superpixel patches using the initial geometry. With an iterative update procedure, an energy objective is pursued, where each superpixel is assigned to the region that satisfies the minimum energy cost. Region updates are terminated either after a fixed amount of iteration or if the energy reduction after the update is smaller than a threshold.

2.1. Superpixel Extraction

The pixel representation of an image is often redundant due to the spatial similarity in the image. In order to reduce this redundancy, a preprocessing stage is introduced by Ren and Malik [9]. This method groups pixels into homogeneous image regions; called superpixels (SPs). 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. Inspired by this idea, all the pixels in a SP can be assigned to specific models representing motion, depth or segmentation structures. This replaces the use of pixels in a graph based representations with SPs. Moreover, when the graph nodes are constructed with SPs instead of pixels, graph complexity and computation time is substantially reduced.

There is a considerable amount of previous work in the literature regarding SP extraction. In our framework, we utilized a method based on [10] due to its fast execution time and convexity property in the extracted SP structure.

2.2. Segmentation with Geometrical Constraints

The proposed region segmentation method includes three main algorithmic steps as shown in Figure 2; 1) Initialization of the regions with the input geometry; 2) Region boundary update; 3) Region structure update.

1) In the first step, generated SPs are assigned to the regions according to the initial geometry. The SP boundaries as shown in Figure 3(a) are initialized to the input $3 \times 1$ geometry in Figure 3(b).

2) The second boundary update step performs a greedy search on the boundary SPs. Figure 3(c) shows the SP updates on the regions boundaries. During the boundary adaptation, the cost function that relates the similarity of the SP to the corresponding region candidates is minimized. This approach assures that the final regions are composed of connected SPs without any sub-detachment. SP to region assignment is performed according to the formula given in (1), where $L(p)$ is the region label of the SP $p$; $S(p, Q_i)$ is the similarity cost between the corresponding SP $p$ and region $Q_i$. $N$ is the number

![Fig. 2. Algorithmic flow of the proposed region segmentation](image)

![Fig. 3. Region segment generation for $3 \times 1$ geometry.](image)
of neighboring region candidates. Therefore, starting from the initial region geometry, boundary SPs are reassigned to the most similar neighboring regions.

\[ L(p) = \arg \max_{i=1:N} \left( S_i(p, Q_i) \right) \] (1)

3) During the structure update, the region statistics are recalculated based on the removed or merged boundary SPs. This update provides SP groups to adapt changes along the region boundaries and converge to compact and coherent region model. The boundary and structure update steps are iterated until the stopping criteria is met. Termination criteria can be set as a fixed number of iteration or it can be computed depending on the decrease in energy cost during the update step. Figure 3(d) presents the generated region segments after the termination condition is met.

The optimization rule given in (1) updates the region boundary. Each boundary SP is visited and assigned to the region that provides maximum similarity. The proposed cost function used is composed of two main energy terms as denoted in (2). The first term relates the color similarity of the boundary SP to its neighboring regions. The second term defines the spatial distance of the SP to the region centers.

\[ E(p, Q) = \lambda \ C(p, Q) + (1 - \lambda) \ D(p, Q^c) \] (2)

The term \( \lambda \) in (2) is a trade off parameter to be tuned depending on the content. Selection of \( \lambda \) imposes the geometry constraint on the generated region segments. As \( \lambda \) is increased, the input geometry constraint will be relaxed in favor of color similarity in the region. In this paper we used 0.5 in all experiments. This value is selected as a mid point between the imposed geometry constraint and region color similarity. Lab color space is utilized in the experiments due to its perceptual uniformity. Color distance is computed over the individual color channels \( i \), see equation (3).

\[ C(p, Q) = \sum_{i=1}^{3} |p_i - Q_i|^2 \] (3)

The spatial distance between the boundary SP \( p \) and the region centroid \( Q^c \) is computed using the geodesic distance. It is defined as the length of the shortest path from \( p \) to \( Q^c \), as given in (4) [11].

\[ D(p, Q^c)_G = \min_{\rho=p_1, p_2, \ldots, p_n} l(P) \] (4)

Suppose \( P = p_1, p_2, \ldots, p_n = Q^c \) is a path between the SPs \( p_i \) and \( p_n = Q^c \) where \( p_i \) and \( p_{i+1} \) are connected neighbors. The path length \( l(P) \), as defined in (5), is the sum of individual neighbor distances \( d_N(p_i, p_{i+1}) \) between adjacent SPs in the path.

\[ l(P) = \sum_{i=1}^{n-1} d_N(p_i, p_{i+1}) \] (5)

For the computation of adjacent SP distance \( d_N \), three color channel (Lab) distance is utilized (6).

\[ d_N(p, q) = \sum_{k=1}^{3} (p_i - q_i)^k \quad k = 1, 2 \] (6)

No significant performance difference has been observed in the selection of \( k \), hence, it is selected as 1 in all the experiments due to its computational efficiency.

Computation of the shortest path from the boundary SP to the region centroid is performed via the shortest path algorithm in [12]. At each iteration, shortest paths from the neighboring boundary SPs to the region centroid are computed. Since the termination criteria for path computation is at the boundary, calculation of the shortest paths over the whole image is avoided.

3. EXPERIMENTS

The advantage of the proposed region segmentation method is validated as an improved pooling concept for image classification. The utilized training based classification method aims to assign the test samples in the dataset to one of the predefined classes. We utilized the PASCAL VOC 2007 [13] image classification dataset for the training and evaluation. This challenging dataset is composed of 9963 images (5011 for training 4952 for testing) and 20 classes including: person, motorbike, airplane, car, cow, bottle, sofa, etc. The measure used to calculate the performance of a given system is the Mean Average Precision (MAP) [14].

A typical image classification pipeline can be explained in the following steps: 1) Extraction of local image features; 2) Encoding of local image descriptors; 3) Pooling of encoded descriptors into a global descriptor; 4) Training and classification of pooled image descriptors.

We use a standard pipeline of process as detailed in [14]. First, the 128-dimensional SIFT [15] descriptors are densely extracted from the entire image. A Principle Component Analysis step is performed in order to reduce redundancy. Dimensionality reduction to 64 has shown considerable improvement in the final classification performance.

For feature encoding, we selected the state-of-the-art Fisher Vectors method [16] for its superior performance in the image classification tasks. It utilizes the first and second order statistics of the difference between the image feature data and the Gaussian mixture model (GMM) learned on the training image descriptors.

The pooling step aims to combine the responses of encoded descriptors spatially or in the feature space. Spatial pyramid [17, 3] is a common pooling strategy that introduce a weak geometry in the encoding phase. Regions in the image are generated as follows: \( 1 \times 1, 1 \times 3 \) (three horizontal stripes), and \( 2 \times 2 \) (four quadrants) grids. Our contribution by the proposed region segmentation is a replacement of this fixed image partitions. We propose that the spatial coherency in the pyramid segments can be better exploited with the proposed geometrically constrained region segmentation.
Number of GMMs

<table>
<thead>
<tr>
<th>Spatial Pyramid Type</th>
<th>Conventional</th>
<th>Proposed</th>
<th>Conventional</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 × 1</td>
<td>52.90</td>
<td>52.90</td>
<td>54.87</td>
<td>54.87</td>
</tr>
<tr>
<td>3 × 1</td>
<td>55.40</td>
<td>55.71</td>
<td>57.09</td>
<td>57.55</td>
</tr>
<tr>
<td>2 × 2</td>
<td>53.71</td>
<td>54.57</td>
<td>55.60</td>
<td>56.98</td>
</tr>
<tr>
<td>1 × 1 + 3 × 1</td>
<td>56.40</td>
<td>56.88</td>
<td>58.17</td>
<td>58.28</td>
</tr>
<tr>
<td>1 × 1 + 2 × 2</td>
<td>55.28</td>
<td>56.12</td>
<td>57.02</td>
<td>58.05</td>
</tr>
<tr>
<td>1 × 1 + 3 × 1 + 2 × 2</td>
<td>56.99</td>
<td>57.61</td>
<td>58.71</td>
<td>59.36</td>
</tr>
</tbody>
</table>

Table 1. Pascal VOC classification results (MAP) with different spatial pyramid combinations

In the final classification step, the linear Support Vector Machine (SVM) method has been used. This method has been shown to perform well especially with the higher dimensional Fisher vector encoding. The SVM model is trained independently in a one-to-all fashion for each image class using the training set. The test scores are ranked depending on the output likelihood of each image belonging to one of the 20 classes in the dataset.

4. RESULTS

Mean average precision (MAP) scores for the each 20 class of the dataset are presented in Table 1. The experiments are conducted for two different GMMs sizes (128 and 256). Comparative tests for different geometry assignments and possible pyramid combinations are performed. (1 × 1, 3 × 1, 2 × 2 and possible combinations of these individual geometries). With the repeated experiments, the standard deviation in MAP scores is observed to be less than 0.2%.

A general observation of the MAP scores in Table 1 indicates that spatial pooling with the proposed region segmentation introduces a consistent increase in the classification accuracy for all of the tested scenarios. The resulting MAP scores are observed to be in line with our initial hypothesis that the proposed region segmentation can better encapsulate local coherency compared to the conventional fixed region assignment. However, one can still argue that there is room for improvement with selection of different region geometries.

Moreover, a detailed look at the change in the AP scores of the individual classes for the proposed segmentation might supply more information. Figure 4 shows the difference of the class specific average precision scores for the individual 20 classes. Red corresponds to the change in AP for the 3 × 1 and green for 2 × 2 spatial region geometry. One can observe that in the 3 × 1 geometry, 12 out of 20 classes have benefited form the proposed segmentation. However in the 2 × 2 case, 18 out of 20 classes have been observed to gain accuracy improvement. This result would clearly show the advantage of the proposed method especially for 2 × 2 geometry. This could be reasoned with the conclusion that the 2 × 2 geometry is inadequate for encapsulating region properties for the used

Fig. 4. Change in AP scores for individual classes with the 3 × 1 and 2 × 2 geometry

Pascal VOC Dataset. Therefore, the proposed region adaptation has shown significant improvement for the great majority of the object classes.

5. CONCLUSION

In this paper, we propose a method for adapting spatial pyramid regions using a predefined geometry constraint. Input geometry is enforced in the segmentation for incorporating spatial coherency in the image. This has been done by using a geometry constrained energy function. The idea has been experimentally evaluated on the image classification pipeline as a replacement to the standard spatial pyramid pooling. Mean average precision scores support the proposed hypothesis that coherent spatial regions would consistently improve image classification performance for alternative scenarios. Sample visual results of the proposed segmentation are also supplied for illustration of the region adaptation.

The proposed method is designed to be generic; hence, it could be integrated into any image classification pipelines to improve the accuracy of the conventional spatial pyramid method with minimal extra effort.

Individual class precision scores have shown that 2 × 2 geometry benefits more from the proposed segmentation. This can be related to inadequate region coherency in the 2 × 2 geometry. This indicates a future perspective towards exploration of scene geometry for region segmentation.
6. REFERENCES


