ON DENSE SAMPLING SIZE
Xue Li, Hongxun Yao, Xiaoshuai Sun, Yanhao Zhang
School of Computer Science and Technology, Harbin Institute of Technology, China
{xueli, h.yao, xiaoshuaisun, yhzhang}@hit.edu.cn

ABSTRACT
This paper proposes a general method for size optimization in dense sampling to obtain a better representation of an image. Our method can be utilized to improve the performance of image classification and other tasks. We discuss the spatial consistency in global-scope restrained descriptors, by analyzing the appropriate sampling size. We apply the low rank method to solve the representative matrix of the descriptor sets at different scales, and obtain the optimized dense sampling size according to the lowest ranks of the representative matrices. Experimental results indicate that the proposed method gives an innovative and effective image representation, and it outperforms traditional dense sampling without size optimization.

Index Terms— dense sampling, sampling size, low rank, image classification

1. INTRODUCTION
Image representations play pivotal roles in many computer vision tasks, including image understanding, object recognition and scene classification. Solid local feature extraction provides abundant information and robustness to deformation and occlusions, which is a fundamental topic arising since decades ago. Among them two of the most common techniques are sophisticated interest points (Harris, SIFT, etc.)[1, 2, 3] and dense sampled points[4, 5, 6].

Interest point detectors focus on “interesting” local regions that can help build correspondences between images of the same or similar object or scene. To this end, high repeatability is required, which is guaranteed by the accuracy and reliability of the feature extraction procedure.

Dense sampled descriptors on a regular grid overcome several drawbacks of interest points, which include limited information, limited coverage regions and subjective prior knowledge. Dense sampling of the visual words combined with the image cues has recently demonstrated significant performance in image interpretation, classification or other tasks. The synthetic results of the PASCAL Visual Object Classes Challenge (VOC) 2010 [7] show that Spatial Pyramid with dense SIFT or overlapped HOG performs superbly among the current descriptors.

However, blindly dense sampling strategy results in low repeatability compared with interest points, caused by the inaccuracy of sampling scales and locations. Attempts to resolve this dilemma result in some related works. T. Tuytelaars[8] introduce dense interest points to merge the advantages of both schemes, applying each densely sampled image patches the local optimization of scale and location within a bounded search area. Yousun Kang et al. [9] incorporate scale information into dense visual word textons as local textural context of the object in image categorization and segmentation. Instead of patch-based representation, [10] propose a region-based idea to represent the image as a map of densely overlapping regions, which can fully utilize the connectivity information between regions. D-Nets [11] constructs a network to analyze image content sampled from selected edges in this net.

Research to date has been devoted to explore the proper representation of descriptors. Moreover, in the extraction of traditional descriptors, limited local information is involved, and global information of the whole image is ignored. Descriptors in an image have distribution associations and spatial consistency, which means descriptors in an image tend to share same or similar scales within a certain area, and an inaccurate-size sampling often break their associations and consistency. For this reason, we analyze the optimized size of dense sampling by maximizing the spatial consistency to achieve better feature representation.

We propose the method of searching for a low-dimensional representation of a high-dimensional feature set in different sampling spaces to handle this spatial consistency maximization problem. In recent years, many methods based on sparsity and rank minimization have been proposed. Among them, low-rank representation [12] seeks the representation matrix with the lowest matrix rank by
taking the image as a dictionary and solving the relative convex optimization problems, which is reported to perform promising for both noiseless and noisy cases. Inspired by this, we propose to seek the representation matrix with lowest rank under multiple spatial spaces. The performances of feature descriptors in multiple sampling sizes are evaluated, and the optimized sampling size for each image in a specific class is learned, which is applied as the solution of the size optimization problem testing procedure.

The remainder of the paper is organized as follows: the details of the method are introduced in Section 2. The performance is evaluated experimentally in Section 3. Finally, we draw the conclusion in Section 4.

2. PROPOSED METHOD

2.1. Framework

The framework of the proposed sampling size optimization method is shown in Figure 2. To search for a best representation of an image at multiple spatial spaces and learn the optimized sampling size from the dataset, we extract multiple patch-size dense descriptors (e.g. dense SIFT) to analyze the images.

Since the sizes of the descriptor vectors vary with sampling sizes, a normalization to align the descriptor vectors is processed to avoid their influence on the dimensions of the representative matrices $Z$, and further the ranks. Each of the normalized descriptor sets of different spatial patch sizes is taken as the input of the convex optimization problem. Then the ranks of the representative matrix $Z$ for each descriptor set are estimated by the optimization formula. After statically analyzing the ranks, the optimized sampling size for each of the image descriptor sets can be obtained, as shown in the figure. Finally, the features extracted at their corresponding optimized scales can be applied in the classification and other computer vision tasks.

2.2. Low-Rank Representation

As we put in Section 1, we want to rectify the image representation by maximizing the spatial consistency, then we pose this problem as a method searching for a low-dimensional representation of a high-dimensional feature set in different sampling spaces. The reasons we use low-rank representation for the descriptors in a whole image are as follow,

- Low-rank representation for the entire image indicates a global constraint. Compared with local information only, this constraint matters a lot, since dense sampling on a regular grid is a global action making all of the units sharing the same calculating scales.

- Low-rank representation increases the spatial consistency. As we know, continuity is an important character for images, which is reflected in spatial consistency -- same/similar features and scales within a certain area. In contrast to sparse interest points, descriptors of dense sampling face this spatial consistency problem because of their continuous locations and strong coverage. Low-rank representation can maximize the spatial consistency by seeking for a low-dimension representation.

- Further experiments demonstrate a substantial correlation between the low-rank mechanism and the dense sampling representation of image.

Let $D_k = \{d_{i1}, d_{i2}, \ldots, d_{in}\}$ be a descriptor set with $n$ descriptors $d_k$ (such as dense SIFT descriptor) extracted at size $k$. The numbers of the dense descriptors $n$ for various sampling sizes are different. When an image is divided into $a \times b$ patches (where $a$ and $b$ are the cell numbers for rows and columns), the descriptor number for this image is $n = a \times b$. So with the larger cell number, we get smaller sampling size, and more descriptors. Each of the descriptor sets can be represented by the product of the dictionary $A$ and a coefficient matrix $Z$:

$$D_k = A_k Z_k$$

where $A_k = \{a_1, a_2, \ldots, a_n\}$ and $Z_k = \{z_1, z_2, \ldots, z_n\}$, subscript $k$ is the corresponding spatial sampling size and $Z_k$ is the representation matrix of $d_k$. According to low-rank representation [12], since each descriptor of the patches is independent from each other, to compute the pairwise affinities to form the affinity matrix, we use the descriptor set itself to be the dictionary.

$$D_k = D_k Z_k$$

$Z$ can be at least one block diagonal matrix, and there are actually infinite feasible solutions to Eq-2.
To normalize the varied descriptor numbers $n$ for the descriptors in different sampling sizes, we randomly selected the same number $M$ of descriptors for all the different sampling strategies. Let randomly selection formula is $\text{random}(D_k, M) = r_{ik}(D_k)$, then Eq-2 can be written as

$$r_{ik}(D_k) = r_{ik}(D_k)Z_{ik}$$

To seek the representation matrix at the optimized scale with the lowest rank, the optimization problem with sampling size information added is formulated as:

$$\begin{align*}
\min & \| Z_{ik} \| + \lambda \| E \|_1, \\
\text{s.t.} & r_{ik}(D_k) = r_{ik}(D_k)Z_{ik} + E
\end{align*}$$

(4)

$$\| E \|_1 = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m} E_{ij}^2}$$

(5)

where $\| \cdot \|$ is the nuclear norm of a matrix. $E$ is the matrix of errors representing the noise to the key elements of the descriptors. Parameter $\lambda > 0$ is the weight to balance the clean part and the corrupted one. Since the images involve complex and compositive objects, the weight $\lambda$ is set to a relatively large number.

This optimization problem can be solved by inexact Augmented Lagrange Multiplier (ALM) algorithm [13]. Therefore the representative matrices of the descriptors with the lowest rank at multiple sampling sizes can be obtained, to be analyzed in the optimized size selection session.

2.3. Optimized Sampling-Size Selection

From the aspect of the spatial consistency of the descriptors, the representative matrix with lowest rank can be understood as the contraction result of the image space. Therefore the rank of the corresponding representative matrix shows the degree of consistency. That means the comparison between the lowest ranks of the descriptors sampled from the same image at different spatial scales can express the difference degrees of spatial consistency, which is further reflected in the performance of key feature expression and image classification.

The lowest rank of the representative matrix is defined as the sum of the singular values:

$$Z_k = USV^T, \quad S = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$$

(6)

$$r_k(Z) = \sum_{i=1}^{n} \sigma_i$$

(7)

where for $n \times m$ matrix $Z$, there are matrix $U$ and $V$ rank $n$ and $m$.

For different spatial scales, the representative matrix with a low rank stands for a high level of spatial consistency, thus the probability of a scale to represent the features of an image is defined as:

$$P_{\text{scale}}(i) = \frac{1}{r_k(Z), i = 1, 2, \ldots, k}$$

(8)

where $i$ is the sampling size index. Therefore the optimized size is the one with the largest probability. When the ranks of two representative matrices are the same, we choose the one with smaller size and more cells because of the abundant information. Then the selected reintegration descriptor set for each image is

$$D_{\text{opt}} = \begin{cases} 
D_{jr}, P_{\text{scale}}(j') > P_{\text{scale}}(i), i = 1, 2, \ldots, k \\
D_{jr}, P_{\text{scale}}(j') = P_{\text{scale}}(j), i < j \neq 1, 2, \ldots, k 
\end{cases}$$

(9)

where each feature of the descriptor set $D_{jr}$ is dynamically learned from the size probability.

Based on the reintegration descriptor set with effective sampling information, we extract bag-of-words features (1024 clusters), and then apply linear classifier to get the final classification result.

3. EXPERIMENTAL EVALUATION

3.1. Experimental Dataset and Settings

The proposed approach was evaluated on the 15-class scene dataset, which was gradually built by Oliva and Torralba [14], Fei-Fei and Perona [15] and Lazebnik et al.[16]. The dataset consists of 4485 images of 15 categories, where each category contains 210 to 410 images, of about 250*300 resolution. The categories of scene-15 range from natural scenes like forests and mountains to man-made environments like bedrooms and offices.

We choose 70 percentage of each category as the training data, and the rest for testing. Dense SIFT descriptors on regular grids are extracted in each image, by dividing them into equal $n \times n$ cells, from $15 \times 15$, $17 \times 17$, to $45 \times 45$. We have a total of thirteen sampling sizes, where cell number $r = 15, 17^2, \ldots, 45^2$. There is no grid spacing between cells.

3.2. Scene Classification

We evaluate the performance of the low-rank densely sampling strategy in the scene classification task. Figure 3 shows the ranks of the representative matrices for four images in different cell numbers. The bigger cell number one has the smaller the sampling size, and the more the descriptors. Generally speaking, higher ranks indicate the complexity of images. For example, image 2897&1255 hold more complicated to be represented. For each image, the sampling size hits the lowest rank is the optimized solution. Image No. 2897 reaches the lowest rank in bigger sampling sizes than the other three images, which means it fits a bigger sampling size.
We count the numbers of cells hits the minimum rank for all the images in different sampling procedures. The boxplot for the statistics is shown in Figure 4. The six variables indicate upper quartile, most frequent position, sample maximum, sample minimum, and lower quartile. After studying the symmetry of the boxplot (‘most’ value is excluded), we can see that there is little skew, which means a high level of dispersion and few outliers in our experiments. The high dispersion results from the widely distributed images with wide-ranging scale by sampling in different cell sizes. The most frequent cell numbers are also added as the ‘most’ value, which means the optimized sizes are different in each of the categories, and it presents our method can discriminate each of the classifications well.

We compare our proposed method with basic dense sampling. For each category, we average the classification accuracy under different sampling size (Table 1 DS(average)), and select the best sampling size for images within each category(Table 1 DS(best)). The performance of our proposed method, which provides an optimized sampling size for each image, is shown in Table 1 as well. Cell# stands for the cell number when the DS performs best. Our proposed method surpasses the dense sampling (average) for all classes, and outperforms dense sampling (best) for 9 out of the 15 classes. Note that the best performances for dense sampling in all the classifications cannot be achieved at the same time, since the best cell numbers and the sampling sizes are different.

Table 1  Results on the Scene-15 classification among dense sampling (average for each class), dense sampling (best performance for each class), and our proposed method.

<table>
<thead>
<tr>
<th>Category</th>
<th>DS(average)</th>
<th>DS(best)</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>suburb</td>
<td>0.84</td>
<td>0.92(32)</td>
<td>0.85</td>
</tr>
<tr>
<td>coast</td>
<td>0.77</td>
<td>0.79(25)</td>
<td>0.84</td>
</tr>
<tr>
<td>forest</td>
<td>0.91</td>
<td>0.93(40,35,42)</td>
<td>0.94</td>
</tr>
<tr>
<td>highway</td>
<td>0.68</td>
<td>0.73(42)</td>
<td>0.84</td>
</tr>
<tr>
<td>inside city</td>
<td>0.64</td>
<td>0.75(17,45)</td>
<td>0.76</td>
</tr>
<tr>
<td>mountain</td>
<td>0.57</td>
<td>0.67(40)</td>
<td>0.69</td>
</tr>
<tr>
<td>open country</td>
<td>0.50</td>
<td>0.56(27)</td>
<td>0.55</td>
</tr>
<tr>
<td>street</td>
<td>0.73</td>
<td>0.79(45)</td>
<td>0.82</td>
</tr>
<tr>
<td>tall building</td>
<td>0.69</td>
<td>0.75(40)</td>
<td>0.71</td>
</tr>
<tr>
<td>office</td>
<td>0.70</td>
<td>0.77(27)</td>
<td>0.72</td>
</tr>
<tr>
<td>bedroom</td>
<td>0.43</td>
<td>0.55(20)</td>
<td>0.63</td>
</tr>
<tr>
<td>industrial</td>
<td>0.31</td>
<td>0.39(45)</td>
<td>0.35</td>
</tr>
<tr>
<td>kitchen</td>
<td>0.44</td>
<td>0.48(27)</td>
<td>0.66</td>
</tr>
<tr>
<td>living room</td>
<td>0.48</td>
<td>0.54(40)</td>
<td>0.59</td>
</tr>
<tr>
<td>store</td>
<td>0.55</td>
<td>0.70(45)</td>
<td>0.63</td>
</tr>
<tr>
<td>average</td>
<td>0.62</td>
<td>0.65</td>
<td>0.70</td>
</tr>
</tbody>
</table>

4. CONCLUSION

In this paper, we introduce a general densely sampling size optimization method to obtain a better expression of an image, and improve the performance of image classification and recognition with the optimized size. With the discussion of the spatial consistency in global-scoped descriptors, we apply low rank representation to achieve the optimized dense sampling size, which leads to lowest rank of representative matrix of descriptors.

The experimental results indicate that our proposed method can provide an innovative and effective image representation. With the optimized size, our method improves the performance of dense sampling. This approach can be applied in all kinds of dense sampling processing as an informational constraint for descriptors. It should be noted that the optimization procedure is time consuming, which we will focus on to improve our method in future work.

5. ACKNOWLEDGMENTS

This research is supported by the National Natural Science Foundation of China (Grant No. 61071180) and Key Program (Grant No. 61133003).
5. REFERENCES


