Visual Words Sequence Alignment for Image Classification

Paweł Drozda
Department of Mathematics and Computer Sciences,
University of Warmia and Mazury,
Olsztyn, Poland
Email: pdrozda@matman.uwm.edu.pl

Przemysław Górecki
Department of Mathematics and Computer Sciences,
University of Warmia and Mazury,
Olsztyn, Poland
Email: pgorecki@matman.uwm.edu.pl

Krzysztof Sopyła
Department of Mathematics and Computer Sciences,
University of Warmia and Mazury,
Olsztyn, Poland
Email: ksopyla@uwm.edu.pl

Piotr Artiemjew
Department of Mathematics and Computer Sciences,
University of Warmia and Mazury,
Olsztyn, Poland
Email: artem@matman.uwm.edu.pl

Abstract—In recent years, the field of image processing has been gaining a growing interest in many scientific domains. In this paper, the attention is focused on one of the fundamental image processing problems, that is image classification. In particular, the novel approach of bridging content based image retrieval and sequence alignment domains was introduced. For this purpose, the dense version of the SIFT key point descriptor, k-means for visual dictionary construction and the Needleman-Wunsch method for sequence alignment were implemented. The performed experiments, which evaluated the classification accuracy, showed the great potential of the proposed solution indicating new directions for development of new image classification algorithms.

I. INTRODUCTION

CBIR domain has recently been of major importance in many fields of everyday life. It has been widely studied and used in such areas as radiology [1], microscopy [18], astronomy [5], mineralogy [19] or e-commerce [15]. For example, in radiology the search of visually similar parts of medical images is performed in order to facilitate the assignment to a particular category of diseases. Among the main tasks of CBIR domain, this paper focuses on the classification. This method assigns each of the queried images to one of the predefined categories, which contains the most visually similar images.

The quality of the results obtained during the classification process largely depends on the features that will be taken into account during comparisons. In the literature, many different ways were proposed to visually describe any image. In particular, color and texture can represent the image globally while for local description the visual keypoints are introduced. On the basis of these characteristics the feature descriptors are formed [21]. For purposes of this paper, the SIFT [16] algorithm was chosen as it is one of the best state of the art image descriptors. On the basis of local image descriptors the different image representations can be considered, which are frequently the vector with a predefined number of features. One of the most commonly used for classification purposes is the Bag of Visual Words (BoVW) image representation, which derives from text categorization and encodes the image as the frequencies of visual local descriptors.

In this paper, the main aim is to perform the classification for the e-Commerce domain purposes with the use of the dataset containing the categorized shoe images. In particular, the idea of bridging BoVW image representation with sequence alignment algorithm for determining image similarities is introduced. This allows for the enrichment of the classification process by capturing the spatial relationships in the images and not only the similarity of local and global features.

For this reason, the SIFT descriptors in the structured form are employed. However, each image is represented by the sequence of visual words arranged in a linear order. To determine the similarity between each pair of images the Needleman-Wunsch algorithm was introduced, which is one of the most commonly used sequence alignment methods for finding the best global matching of two sequences.

The idea of bridging the sequence alignment and CBIR was constituted in Hung-sik et al [12], [13], where the authors use the biological methods for determining image similarities. This paper extends this approach by introducing the BoVW image representation and loosening the biological restrictions in sequence alignment. The performed experiments proved the correctness of the presented method and indicated new potential solutions to the image classification problem.

The rest of this paper is organized as follows. The next section describes existing approaches in the Content Based Image Retrieval field which use the spatial information. In
section III we propose the novel method for image classification, where the sequence alignment and CBIR are merged. Then, the accuracy results of classification are evaluated in the experimental section. Finally, in section V we conclude the paper and point out future work.

II. RELATED WORK

The idea of the Bag of Visual Words (BoVW) approach has its origins in the Information Retrieval discipline. The first experiments with BoVW mimic the procedure of obtaining a numerical representation derived from the IR field. The most computationally intensive and crucial part of this procedure is extracting the visual words, which consists in detecting the interest points on pictures, forming the local descriptors around interest points and quantizing them into visual words. In the literature many detectors and descriptors were proposed like ASIFT [17], [26], SIFT [16], SURF [4], MSER [8], each method tries to be as invariant to translation, rotation, scale changing, illumination and view point as possible. Simple histograms of quantized descriptors works quite well in many CBIR tasks, but pictures has far richer spatial structure than text and exploiting this additional information its beneficial for different image understanding tasks.

The first attempt to use spatial information can be seen in Lazebnik et al [14]. Their spatial pyramid matching algorithm was motivated by Grauman et al. [10] pyramid matching which was designed for finding correspondence between sets of high-dimensional points. Lazebnik uses this idea to find correspondence between sets of visual words, by repeatedly subdividing the image and computing histograms of local features at increasingly coarse resolutions. Utilization of SIFT descriptors on a dense grid with the spatial pyramid matching kernel improves classification accuracy on Caltech 101 dataset. Saverse et al [20] try to exploit the relative spatial locations and build the correlograms of visual words like Huang et al [11] use correlograms of quantized colors. Yang et al [24], [25] go a step further and instead of using the absolute spatial location, they use the relative spatial co-occurrence. Firstly, the particular image is partitioned into a sequence of increasingly coarser grids and in each grid cell the co-occurrence matrix is computed. To get the final similarity score between two images the weighted histogram intersection kernel was used.

Another approach utilizing spatial relations for image matching was suggested by Hung-sik et al [12], [13]. In their work the image is represented as a DNA or protein sequence converted from features extracted from colors and texture. Matching was implemented with use of the BLAST [2] sequence alignment algorithm. Their solution suffers from two limitations. The first one is connected with the utilized WU-BLAST implementation, which works only with 4(DNA) or 23(Protein) element alphabets. In order to create such an alphabet they had to perform translation from the feature space to the particular alphabet. To perform such a conversion the “Composite Conversion Table” was developed, but it is hard to create universal mapping for different features. The second issue was connected with the substitution matrix, which determines the similarity between two alphabet letters and penalty when one is substituted by the other. They use a simple Uniform matrix (1 in diagonal, -1 for all other elements) which means that letters are only similar to themselves and the Gaussian distributed matrix which tries to express similarity between the mapped features, for example color red is more similar to orange than blue.

Our work improves the solution proposed by Hung-sik by using the BoVW approach which in our opinion is more suitable for sequence alignment framework. Firstly, it is more natural to obtain the substitution matrix by finding the distance between each pair of vocabulary centroids, secondly our solution is not limited to particular alphabet lengths. All the implementation details are described in section III.

III. METHODOLOGY

This section precisely describes the whole image classification process proposed in this paper. In particular, two main phases can be pointed out. The first involves creating the sequence image representation with the use of dense SIFT descriptors and the second incorporates two different methods for image classification basing on the similarities of images from the considered dataset. The similarities are obtained by means of the Needleman-Wunsch sequence alignment algorithm.

A. Image as a Sequence of Visual Words

In this study image is represented as a sequence of letters from predefined alphabet, in this context each visual word is treated as a letter. In order to obtain such representation the standard methodology of creating Bag of Visual Words approach was used.

To describe local image patches the SIFT algorithm, was chosen since it was proven to be resistant to scale, rotation and viewpoint [16], [21], [23]. In particular, the SIFT descriptor on the dense regular mesh with the predetermined grid size (8, 16, 24, 32) was applied for the images from the classified dataset, the DSIFT implementation from VLFeat library [22] was utilized. It should be noticed, that all images were scaled to the same size (wider edge equals 256px) in order to provide the same number of grid cells, this makes the classification process more reliable. As a result, for each image the set of the SIFT descriptors (128 dimensional vector enriched with the information about scale, orientation and location) was collected. An example of an image with dense SIFT descriptors is presented in Figure 1(b).

On the basis of all the obtained descriptors in the whole dataset the visual dictionary is created. The dictionary is generated with the \(k\)-means algorithm, where each cluster is a visual word of the generated dictionary. The described process is performed, since the obtained SIFT descriptors are similar and it is worth to group them into clusters. Next step of the algorithm involves assigning the visual word to each cell descriptor, in such a way that the particular cell will contain the most similar visual word to its descriptor. At this stage the visual word can be treated as a single letter. Having obtained
the visual word grid representation of the image (illustrated in Figure 1(c)), the different ordering can be applied to obtain the image sequence representation, needed for Needleman-Wunsch algorithm. In particular, the visual word can be set in the linear order, where grid cells are ordered row by row or column by column, the Hilbert order, where cells are addressed according to Hilbert space-filling curve or Z-order (Morton order). For purpose of this study only the linear ordering is taken under consideration.

B. Sequence Alignment for Classification

The first phase of the algorithm generates the sequences of visual words for all images from the given dataset which will be processed in the second phase by the Needleman-Wunsch sequence alignment algorithm in order to determine the similarities of any pair of images. This method was originally designed for bioinformatic tasks such as aligning protein or nucleotide sequences. It finds the best possible global matching between two sequences $A = A_1 A_2 \ldots A_n$ and $B = B_1 B_2 \ldots B_m$, where $A_i$ and $B_j$ are the letters from the predefined alphabet $W$. In our case the alphabet is the visual dictionary generated in the first phase, which consists of visual words.

Besides of two sequences of visual words for images, the Needleman-Wunsch algorithm takes as an input the scoring, substitution matrix $S(W_i, W_j) = \{s_{ij}\}$. Each value of the matrix indicates the similarities between two alphabet letters. The higher value of the $\{s_{ij}\}$ means that words $i$ and $j$ are more similar and the substitution of that words during alignment entails less penalty. In bioinformatics domain the creation of the substitution matrix is considered as a hard problem and as yet there has not been proposed any optimal matrix for the protein and DNA sequence alignment. There has been proposed many different types of matrices from which the BLOSUM [7] and PAM [6] matrix should be mentioned as a leading. However, in the case of the visual sequence alignment, we construct the similarity matrix on the basis of the distance between two given visual words (precisely, between the clusters from the k-means algorithm). This allows for capturing the correlation between visual words which in case of the standard BoVW approach becomes difficult. In addition, the elements of the similarity matrix $S$ created for the purposes of the Needleman-Wunsch algorithm are calculated by the following formula:

$$S(W_i, W_j) = (-2) * \frac{d(W_i, W_j)}{\max(d(W_i, W_j))} + 1,$$

where $d(W_i, W_j)$ is the distance between two visual words $i$ and $j$.

The idea of Needleman-Wunsch method can be illustrated by the recursion algorithm 1.

In this algorithm $p$ is the penalty parameter, which decreases the total similarity, when instead of matching there occurs a
gap in one of the sequences. The value of this parameter is set for each case individually and in this paper several different values were tested. The output similarity value for any given pair of sequences is saved in $F_{nm}$ where $n$ and $m$ are the lengths of sequences. As the obtained similarities strongly depend on the length of sequences, they should be normalized in order to weaken the correlation. The normalization is performed by the following formula:

$$ Sim_{A,B} = \frac{Sim_{A,B}}{\text{length}(A) + \text{length}(B)}, \quad (2) $$

where $Sim_{A,B}$ is the similarity score obtained from the Needleman-Wunsch algorithm for sequences $A$ and $B$. As a result of the described procedure the matrix of normalized similarities of all possible pairs of sequences is obtained.

As the last step of the proposed method the classification is performed. For this purpose two different algorithms were implemented. Firstly, the standard $k$-NN algorithm was used with 5-fold cross-validation. The second method for each image finds average similarities for each class and classifies it to the class with the highest average.

The results of classification are described in details in the experimental section.

IV. EXPERIMENTS

The goal of the experimental session was to verify the effectiveness of the proposed methodology for image classification. In particular, we wanted to evaluate the method proposed in this paper over the dataset containing the visually similar categories.

For this reason, the dataset consisted of 200 images of shoes, assigned to five distinctive visual categories was proposed. In particular, each category had 59, 20, 34, 29, and 58 images. Sample images from each of the categories are presented in Figure 2.

To investigate the importance of the grid size selection on classification accuracy, dense sampling over the grid patches of size 8, 16, 24 and 32 pixels were considered. It should be noted that the choice of grid size significantly affect the length of the visual word sequence for each image which apart from accuracy determines the classification time. Intuitively, a smaller grid size should increase the level of accuracy but also it will be extended in time. Successively, densely sampled SIFT descriptors were quantized into visual dictionaries of size $k = \{5,10,20,50,100,200,500,1000\}$, which was to determine the optimal size of the visual dictionary for the classification process. The penalty parameter of Needleman-Wunsch algorithm is identified as the next factor which can greatly affect the level of accuracy. In particular, our method was evaluated for various values of the penalty parameter $p = \{0.01,0.1,0.2,0.4,0.5\}$. Since our approach relies on the $k$-NN classifier, we calculate classification accuracy for 1, 3 and 5 neighbors. Tables I, II and III present the results we have obtained using the $k$-NN classifiers with 5-fold cross validation. Precisely, for each value of the parameter $k = 1,2,3$ only the best set of results for chosen penalty parameter is presented. Although, in the case of different choices of penalty parameter, the accuracy changes slightly (up to 3 percent). It proves as well as the data presented in the tables I, II, III that the selection of both the $k$-NN classifier and penalty parameter has a very little impact on the accuracy of classification. On the other hand, it can be noted that the accuracy is very high for very dense grids (grid size of 8) and dictionaries with more than 20 visual words and it significantly decreases with increasing the grid size. The best reported accuracy is equal to 100% for 1-NN classifier with 200 or more visual dictionary size and in many other cases the accuracy is 98% or more. Our previous experiments on the same dataset [3], [9], which used classical visual word frequencies, $k$-NN

```
Algorithm 1 Needleman-Wunsch algorithm
1: $F_{i0} = p \cdot i, \ i \in \{0, ..., n\}, j = 0$
2: $F_{0j} = p \cdot j, \ i = 0, \ j \in \{0, ..., n\}$
3: $F_{ij} = \max(F_{i-1,j-1} + S(A_i, B_j), F_{i-1,j} + p, F_{i,j-1} + p)$.
```

### Table I

<table>
<thead>
<tr>
<th>Vocab. size</th>
<th>Grid size</th>
<th>8</th>
<th>16</th>
<th>24</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.97</td>
<td>0.9</td>
<td>0.84</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.975</td>
<td>0.92</td>
<td>0.86</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.995</td>
<td>0.945</td>
<td>0.875</td>
<td>0.875</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.985</td>
<td>0.96</td>
<td>0.915</td>
<td>0.815</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1.00</td>
<td>0.96</td>
<td>0.895</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>1</td>
<td>0.955</td>
<td>0.945</td>
<td>0.915</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>1</td>
<td>0.975</td>
<td>0.925</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>1</td>
<td>0.98</td>
<td>0.93</td>
<td>0.905</td>
<td></td>
</tr>
</tbody>
</table>

### Table II

<table>
<thead>
<tr>
<th>Vocab. size</th>
<th>Grid size</th>
<th>8</th>
<th>16</th>
<th>24</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.96</td>
<td>0.885</td>
<td>0.845</td>
<td>0.815</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.97</td>
<td>0.91</td>
<td>0.825</td>
<td>0.875</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.99</td>
<td>0.935</td>
<td>0.9</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.98</td>
<td>0.96</td>
<td>0.92</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.99</td>
<td>0.955</td>
<td>0.895</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.99</td>
<td>0.98</td>
<td>0.91</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.99</td>
<td>0.97</td>
<td>0.93</td>
<td>0.905</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.99</td>
<td>0.975</td>
<td>0.94</td>
<td>0.925</td>
<td></td>
</tr>
</tbody>
</table>

### Table III

<table>
<thead>
<tr>
<th>Vocab. size</th>
<th>Grid size</th>
<th>8</th>
<th>16</th>
<th>24</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.965</td>
<td>0.885</td>
<td>0.82</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.97</td>
<td>0.905</td>
<td>0.86</td>
<td>0.845</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.985</td>
<td>0.94</td>
<td>0.88</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.97</td>
<td>0.935</td>
<td>0.895</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.975</td>
<td>0.945</td>
<td>0.89</td>
<td>0.905</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.96</td>
<td>0.955</td>
<td>0.91</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.99</td>
<td>0.95</td>
<td>0.915</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.99</td>
<td>0.96</td>
<td>0.91</td>
<td>0.91</td>
<td></td>
</tr>
</tbody>
</table>

### Classification accuracy of 1-NN classifier (penalty parameter = .5)

### Classification accuracy of 3-NN classifier (penalty parameter = .01)

### Classification accuracy of 5-NN classifier (penalty parameter = .2)
and SVM classifiers, resulted in a 96% accuracy. Therefore, the classification based on sequences of visual words turns out to be more reliable than classical approaches. The sample results calculated from the second method, based on average similarities for each class, are presented in table IV. In this case the accuracy is lower (the best result is 95%) compared to \( k \)-NN classifiers.

The grid size parameter has the biggest impact on the classification accuracy. The classification accuracy is very high for small grid patches and it degrades if patches get larger. However, the denser grid results in the longer visual sequences of the images. An average sequence length for grid size 8 is 497.1 (\( \sigma = 158.12 \)), while it is 18.2 (\( \sigma = 12.3 \)) for the grid size of 32. As Needleman-Wunsch algorithm requires \( O(mn) \) time to calculate the similarity of two images with sequences of length \( m \) and \( n \), high classification accuracy comes at the price of high computational cost. For example it takes roughly 746x less time to calculate the similarity of two images at the grid size of 32 than 8.

V. CONCLUSION AND FUTURE WORK

This paper presents the novel approach of the image classification by usage of the Needleman-Wunsch sequence alignment algorithm for image similarity finding. The accuracy obtained during the experimental session reaches 100% for properly chosen parameters, which is better than the standard SVM classifier, where the highest accuracy obtained in SVM classification is equal to 96.5% for the same dataset. The experimental results showed the great potential of combining these two research areas and point the new directions in the Content Based Image Retrieval Field.

In the future we are going to test our approach on the state of the art imaginary datasets as well as implement different sequence alignment algorithms. Moreover, we plan to test different methods for sequence ordering.

ACKNOWLEDGMENTS.

The research has been supported by grant N N516 480940 from The National Science Center of the Republic of Poland and by grant 1309-802 from Ministry of Science and Higher Education of the Republic of Poland. We would like to thank Geox and Salomon companies for providing a set of shoes images for scientific purposes.

REFERENCES


