Compact and discriminative approach for encoding spatial-relationship of visual words

Symposium on Applied Computing, 30th, 2015, Salamanca.
http://www.producao.usp.br/handle/BDPI/49013

Downloaded from: Biblioteca Digital da Produção Intelectual - BDPI, Universidade de São Paulo
Compact and Discriminative Approach for Encoding Spatial-Relationship of Visual Words

Glauco V. Pedrosa and Agma J. M. Traina
Instituto de Ciências Matemáticas e de Computação - ICMC
University of Sao Paulo (USP), São Carlos, SP - Brazil
{gpedrosa,agma}@icmc.usp.br

ABSTRACT

The Bag-of-Visual-Words (BoVW) approach has been successfully used for video and image analysis by encoding local features as visual words, and the final representation is a histogram of the visual words detected in the image. One limitation of this approach relies on its inability of encoding spatial distribution of the visual words within an image, which is important for similarity measurement between images. In this paper, we present a novel technique to incorporate spatial information, called Global Spatial Arrangement (GSA). The idea is to split the image space into quadrants using each detected point as origin. To ensure rotation invariance, we use the information of the gradient of each detected point to define each quarter of the quadrant. The final representation uses only two extra information into the final feature vector to encode the spatial arrangement of visual words, with the advantage of being invariant to rotation. We performed representative experimental evaluations using several public datasets. Compared to other techniques, such as the Spatial Pyramid (SP), the proposed method needs 90% less information to encode spatial information of visual words. The results in image retrieval and classification demonstrated that our proposed approach improved the retrieval accuracy compared to other traditional techniques, while being the most compact descriptor.

Categories and Subject Descriptors

I.4 [IMAGE PROCESSING AND COMPUTER VISION]: I.4.10—Image Representation

Keywords

image representation, local features, bag-of-features, spatial-relationship, visual words, visual dictionaries.

1. INTRODUCTION

The Bag-of-Visual-Words (BoVW) approach has been successfully used for video and image analysis by encoding local features as visual words, and the final representation is a histogram of the visual words detected in the image. However, this traditional approach disregards all information about the spatial relationships of the features, affecting the accuracy of recognition. The spatial information is critical for image and object characterization, once varying the image position/placement results in different features. Thus, the potential power of encoding the spatial relationship of visual words has motivated the creation of approaches to tackle this problem [1, 3, 4, 6].

One of the first works proposed to attempt encoding spatial information with BoVW is the “Spatial Pyramid” (SP) [3]. This technique subdivides the image into hierarchical cells and computes a Bag-of-Visual-Words for each cell, concatenating all the histograms generated at the end. The results of this approach showed considerably improved performance in image classification [1, 4]. However, the main drawback of the SP is the large feature vector generated, making its use infeasible in applications with large datasets. A recent spatial coding technique, called “Words Spatial Arrangement” (WSA) [6], uses a different strategy. It computes the spatial relationship of each visual words in relation to all detected points. The main benefit of WSA compared to SP is to provide an adequate performance with much more compact features. However, it is crucial that the image characterization does not depend on the placement of the image, because the features should be invariant to geometric transformations.

In this paper, we present a simple but effective approach to encode the spatial distribution of visual words, namely “Global Spatial Arrangement” (GSA). The idea of the GSA approach is to compute the occurrences of the visual words in relation to the other visual words positions, following the strategy adopted by the WSA approach. However, our technique uses the information of the gradient direction to make the representation robust to geometric transformations, such as rotation. Besides that, it uses only two information (top and left) for encoding the visual words arrangement. Due to this fact, the proposed spatial descriptor is more compact and suitable to work with large datasets.

Experimental tests were performed using several different image datasets. Comparative evaluations demonstrate that our proposed spatial descriptor overcomes the SP and WSA approaches, being the best option of spatial descriptor in each evaluated dataset.

2. THE PROPOSED METHOD

The proposed method, called Global Spatial Arrangement (GSA), aims at encoding spatial arrangement of visual words using each local point detected as the origin of a quadrant and counting the number of visual words that appear in each quarter of this quadrant. In other words, for each point $p_j = (x_j, y_j)$ detected in the image, the image space is divided into quadrants with each point $p_j$ as the origin of the quadrants. Then, we count the number of visual words that appears in each quarter of the quadrant. This
procedure is computed for the other points in the image and the counting is incremented. Figure 1 shows an example of partitioning the image space and counting. The red circles are the detected points, labeled by their associated visual words. The idea is to put a quadrant’s origin in each red circle and counting the visual words that appears in each quadrant. When all points have already been the quadrant’s origin, the counting finishes and it is normalized by the sum.

Formally, the first order derivative of an image components given by the derivatives in the horizontal and vertical quadrant. Mathematically, the gradient is a 2D vector with its about the gradient direction of each visual word to define the the image is rotated. To solve this problem we use the information that appears in each quarter of the quadrant. The final counter each detected point and count the number of visual word that points with its visual word. We put the quadrant’s origin in the sum.

\[
I_x(x, y) = I(x + 1, y) - I(x, y)
\]

\[
I_y(x, y) = I(x, y + 1) - I(x, y)
\]

The gradient direction is invariant to rotation, because it always points in the direction of higher values. Due to this fact, this information is used in the proposed approach to define which quarter of a quadrant a point will belong to. Thus, we proposed the following definition to decide which quarter of a quadrant \( Q = (Q_1, Q_2, Q_3 \text{ or } Q_4) \) a point \( p_j = (x_j, y_j) \) will belong considering the gradient direction of the origin point \( p_i = (x_i, y_i) \).

**Definition 1.** Let \( Q \) be a quadrant with the point \( p_i = (x_i, y_i) \) as the origin. We defined the following conditions to decide which quarter (\( Q_1, Q_2, Q_3 \text{ or } Q_4 \)) of the quadrant \( Q \) a point \( p_j = (x_j, y_j) \) will belong to:

- If \( I_x(p_i) > 0 \) and \( I_y(p_i) > 0 \), then
  - if \( x_j < x_i \) and \( y_j < y_i \), then \( p_j \rightarrow Q_1 \)
  - if \( x_j < x_i \) and \( y_j > y_i \), then \( p_j \rightarrow Q_2 \)
  - if \( x_j > x_i \) and \( y_j < y_i \), then \( p_j \rightarrow Q_3 \)
  - if \( x_j > x_i \) and \( y_j > y_i \), then \( p_j \rightarrow Q_4 \)

- If \( I_x(p_i) > 0 \) and \( I_y(p_i) < 0 \), then
  - if \( x_j < x_i \) and \( y_j > y_i \), then \( p_j \rightarrow Q_1 \)
  - if \( x_j < x_i \) and \( y_j < y_i \), then \( p_j \rightarrow Q_2 \)
  - if \( x_j > x_i \) and \( y_j > y_i \), then \( p_j \rightarrow Q_3 \)
  - if \( x_j > x_i \) and \( y_j < y_i \), then \( p_j \rightarrow Q_4 \)

- If \( I_x(p_i) < 0 \) and \( I_y(p_i) > 0 \), then
  - if \( x_j > x_i \) and \( y_j < y_i \), then \( p_j \rightarrow Q_1 \)
  - if \( x_j > x_i \) and \( y_j > y_i \), then \( p_j \rightarrow Q_2 \)
  - if \( x_j < x_i \) and \( y_j < y_i \), then \( p_j \rightarrow Q_3 \)
  - if \( x_j < x_i \) and \( y_j > y_i \), then \( p_j \rightarrow Q_4 \)

- If \( I_x(p_i) < 0 \) and \( I_y(p_i) < 0 \), then
  - if \( x_j > x_i \) and \( y_j > y_i \), then \( p_j \rightarrow Q_1 \)
  - if \( x_j > x_i \) and \( y_j < y_i \), then \( p_j \rightarrow Q_2 \)
  - if \( x_j < x_i \) and \( y_j > y_i \), then \( p_j \rightarrow Q_3 \)
  - if \( x_j < x_i \) and \( y_j < y_i \), then \( p_j \rightarrow Q_4 \)

Using the conditions of Definition 1, the representation will be robust to rotation, because the quadrant is defined by the gradient signal of the origin point. Figure 2 illustrates the Definition 1, where the arrows represents the gradient directions in \( x \) and \( y \). If the gradient signal of \( I_x(p_i) \) has a positive value, then it points upward. If the gradient signal of \( I_y(p_i) \) has a positive value, then it points rightward regarding the origin. This information is useful to define the quadrant in a rotation invariant way.

![Figure 1: Example of the strategy used for encoding the spatial arrangement of visual words. The red circles are the detected points with its visual word. We put the quadrant’s origin in each detected point and count the number of visual word that appears in each quarter of the quadrant. The final counter values are shown in the bottom-right image.](image)

Considering a visual dictionary composed of \( k \) visual words, the dimensionality of the final image feature vector will be \( 5^*k \), being \( k \) for the BoVW histogram and \( 4^*k \) for the spatial arrangement computed by the proposed method. To reduce the feature vector dimensionality and to develop a more compact representation, we
consider each quarter Q1, Q2, Q3 and Q4 in four different groups according to its position: top = \{Q1,Q2\}, down = \{Q3,Q4\}, left = \{Q1,Q3\}, right = \{Q2,Q4\}. The proposed final representation is a 2-tuple \( (S_{t/d}, S_{l/r}) \) considering the relation between top-down and left-right, computed as follows:

\[
S_{t/d} = \frac{\text{top}}{\text{top} + \text{down}}
\]

\[
S_{l/r} = \frac{\text{left}}{\text{left} + \text{right}}
\]

Figure 3 shows each quadrant group. This new representation presents a large gain in precision while reducing the size of the feature vector, as we will show in the next section.

3. EXPERIMENTAL VALIDATION

We present the evaluation of the proposed method for image retrieval and classification using three public image datasets: Corel1000 \(^1\), Texture [2] and 15-Scenes [3]. We used the SIFT descriptor [5] to extract the local image features. For each base, some training images were selected, and their keypoints where detected and represented by the SIFT descriptor. After that, a visual dictionary of 500 visual words were constructed using the k-means clustering algorithm.

We compared our proposed spatial representation (GSA) with: the SP technique [3] and the WSA approach [6]. Table 1 compares the feature vector dimensionality used by the three approaches. Both techniques have a feature vector larger than our approach to encode the spatial information of visual words. Our proposed approach needs 90% less information than SP and 50% less than WSA to encode the spatial information.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Dimensionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP</td>
<td>(20k)</td>
</tr>
<tr>
<td>WSA</td>
<td>(4k)</td>
</tr>
<tr>
<td>GSA</td>
<td>(2k)</td>
</tr>
</tbody>
</table>

For illustration purpose, Figure 4 shows the representation vector of Fig. 4-a and Fig. 4-b using the SP, WSA and GSA to encode the spatial information of their visual words. We can see that, the SP and the WSA approaches present different representation when the image is rotated. Besides that, most values of the SP approach is composed of zeros, it means that precious spaces of the feature vector are not necessary.

For illustration purpose, Figure 4 shows the representation vector of Fig. 4-a and Fig. 4-b using the SP, WSA and GSA to encode the spatial information of their visual words. We can see that, the SP and the WSA approaches present different representation when the image is rotated. Besides that, most values of the SP approach is composed of zeros, it means that precious spaces of the feature vector are not necessary.

Table 1: Feature vector dimensionality comparison, where \(k\) represents the number of visual words in the visual dictionary.

considering the image retrieval evaluation, Fig. 5-a-b show the Precision and Recall curves for Texture and Corel1000 datasets, respectively. For both datasets, the GSA approach boosted the retrieval accuracy compared to the baseline (BoVW) that considers only the histogram of visual words. Our proposed descriptor GSA also presented the best curve among the evaluated spatial descriptor (SP and WSA).

Figure 5: PR curves (a) Texture database (b) Corel1000 database.

For image classification evaluation, we performed experimental tests using the 15-Scenes dataset. For this test we used 100 images of each class to train a classifier and the rest of images were used for test. Table 2 shows the classification rate for each class of the 15-scenes dataset. In this evaluation, the SP approach presented 1% better performance than the proposed method in the overall results. However, considering some individuals classes, such as the “forest” class, the proposed technique improved the classification rate in 8.5%. It also presented the best classification for three other classes (“highway”, “mountain” and “office”) among the evaluated spatial descriptors. Our approach boosted the baseline BoVW in 2.6% considering the overall results.

Considering the cost-benefit, the compact representation and the overall results, the GSA outperformed the evaluated spatial descriptors SP and WSA. It achieves the best results for image retrieval and comparative performance for image classification, with the most compact representation. The experiments show that the proposed approach is a good option in place of the SP and the WSA, because it saves space in a compromise of a good gain in accuracy.

\(^1\)available at: http://wang.ist.psu.edu/docs/related/
4. CONCLUSIONS

In this paper we introduced a novel modeling approach for representing images by taking into consideration the spatial relationships of its visual words. The strategy is to consider each detected point as the origin of a four-quadrant structure. Then, we count the number of visual words that occurs at each quarter of the quadrant. This strategy provides information about the spatial distribution of the visual words in relation to the others. To make the proposed technique invariant to rotation, we used the information of the gradient direction to define each quarter of the quadrant, in a way that if the image is rotated, the descriptor is not affected by this transformation. Besides that, the final description compacts the information of each quarter into just two information elements: we group the quarters in relation to its position (top, down, left and right) and then we compute the relation between top/down and left/right.

The proposed method was evaluated for image retrieval and classification. Experimental results show that the proposed contextual relations between visual words is very useful for boosting the image representation and it achieves significant improvement over the traditional BoVW model.

Acknowledgements

This work is supported, in part, by FAPESP, CAPES, SticAMSUD, the RESCUER project, funded by the European Commission (Grant 614154) and by the Brazilian National Council for Scientific and Technological Development CNPq/MCTI.

5. REFERENCES


