The BOW image retrieval algorithm based on the longest common word string constraint

Mingyu Nie¹*, Jiangzhou Zhang¹, Yang Li¹ and Shuai Wang¹

¹Information Research Institute, Qilu University of Technology (Shandong Academy of Sciences), Jinan, Shandong, 250013, China

*Corresponding author’s e-mail: niemy@sdas.org

Abstract. This paper adds the spatial and semantic relations of the "visual dictionary", the longest common visual word string constraint in the image based on the BoW model. The image is represented at the vocabulary level and the phrase level at the same time. This algorithm can better represent the content of the image, thereby improving the quality of image retrieval. Experiments verify that compared with the BoW model, the accuracy of the BoW image retrieval model based on the longest common word string constraint is well optimized.

1. Introduction

Content-based image retrieval combines computer vision technology, machine learning and information retrieval technology, and has become the mainstream direction of image retrieval research in the past ten years[1]. At present, most cutting-edge image retrieval systems rely on the "Bag of Visual Words" BoW model of images to construct a high-dimensional index structure, and have been successfully applied[2]. However, the disadvantage of the BoW representation model is that in the process of transforming the image into a "visual dictionary", the spatial position relationship of the local features in the original image is ignored. Therefore, studying a higher-level image feature representation model based on BoW optimization, improving the spatial and semantic relationships between "visual dictionaries", better representing the content of images, and improving the quality of image retrieval have become key issues to be solved. This paper adds the longest common visual word string constraint based on the BoW model, and represents images at the vocabulary level and phrase level at the same time, which improves the spatial and semantic relationship between "visual dictionaries" and effectively improves the quality of image retrieval.

2. The BOW model and longest common visual word string constraint

2.1. The BOW Model

The BOW model first appeared in the fields of natural language processing and text information retrieval[3]. The model ignores the grammar, word order and other elements of the text, and treats the document as a collection of several words. The words in the document appear independently, and a set of unordered words (words) are used to express a document.

Unlike text, the content information of an image is continuous. Traditional image features such as colour, texture, and shape describe the entire image on a global level, and cannot assume the role of words in the BOW model[4]. Local invariant features have better distinguishing and descriptive capabilities, have a certain abstract meaning, and can reflect the local feature points of the image to a
certain extent. Features such as SIFT, SURF, etc., with good resistance to light, rotation, and noise[5]. When a target object exists in different images, the local invariant features of this target can be kept unchanged. Similar to text, an image can be regarded as a collection of some local features that have nothing to do with location information. The status of these local features is similar to words in the text, here called "visual words", and the collection of visual words becomes a "visual dictionary".

The visual BOW model algorithm process is roughly divided into four steps. Extract the feature descriptors in the image, generally using SIFT features firstly. Then cluster the similar points of the feature descriptors through a clustering algorithm, and each cluster center represents a visual word[6]. Then the local visual features of the image are mapped to the visual word table and represented by a feature vector. Each dimension of the feature vector corresponds to the sum of the weights of a visual word. Finally, the vector generated by the image is used for image retrieval. The algorithm flow chart is as follows:

2.2. The Longest Common Visual Word String Constraint

The image is projected from a three-dimensional scene to a two-dimensional plane. The positional relationship of the target object is one of the important characteristics of the image. Symbolic projection transforms the symbolic image into a 1-D expression by projecting the symbol to two coordinate axes. In this way, the topological relationship and orientation relationship of the target will be simplified to a 1-D sorting relationship on these two coordinate axes. We analyze the 1-D order relationship of the two coordinate axes by constructing a 2-D symbol string, and the 2-D symbol string can be constructed directly from the symbol graph[7].
As shown in Figure 2 below, the objects in the image are regarded as symbols, and these symbols constitute a string of symbols. Each symbol is equivalent to a target, and the symbol string formed in this way belongs to the symbol string with large granularity[8]. Then we reduce the granularity and use visual words as symbols to construct a small-granularity symbol string. As the smallest description unit that can accurately describe the image, visual words are used as symbols to form visual word strings. Visual word strings can maintain the spatial topological relationship between visual words to a certain extent, which contains a lot of contextual information.

The maximum likelihood criterion is used to measure the similarity of two strings, and the longest common substring obtained is the longest common visual word string[9]. As shown in Figure 2, the longest common visual word string formed between an image and an image can be regarded as a shared implicit mode between the two images. The shared mode between different images is different. Generally, the more similar the images, the more prominent the shared mode, that is, the longer the common visual word string[10].

According to the given method of constructing the longest common visual word string, we can obtain the largest common implicit mode of the two images. This paper defines this implicit mode as the longest common substring of two images converted into a symbol string. It can be seen from the algorithm that this method of constructing visual word strings cannot resist rotation[11]. In order to solve this situation, only the query image needs to be rotated. Specifically, the query image is rotated clockwise by 0, 90, 180, 270, and then the rotated image is used as the query image and the gallery image to construct a visual word string, and the largest of the four is taken as the final result.

3. The BOW image retrieval algorithm based on the longest common word string constraint

This paper innovatively proposes a BoW optimization model based on the longest common word string constraint, combining 2-D string and the longest common word string theory. This model can represent images at the vocabulary level and phrase level at the same time, can more effectively represent the content of the image and the objects it contains, and improve the retrieval quality and scalability of the image. The flowchart is shown in Figure 3.

![Diagram of BoW model based on the longest common word string constraint](image)

Figure 3. The algorithm flowchart of BoW model based on the longest common word string constraint.

The specific algorithm steps of the algorithm are as follows:
Step one: Perform local SIFT feature extraction on the image library to obtain the local feature set of the image library. Clustering the local feature set by K-means clustering method. The cluster centre is the visual word. Build the BOW visual dictionary as \( \text{codebook} = \{w_1, w_2, ..., w_m\} \).

Step two: Extract SIFT features of the query image \( F = \{f_1, f_2, ..., f_L\} \). The feature vector of the query image is assigned to each dimension of the visual dictionary \( \text{codebook} = \{w_1, w_2, ..., w_m\} \) using the nearest neighbour algorithm, and an m-dimensional visual word histogram vector \( H \) is generated. In the process of allocation, each SIFT feature vector in the image is closest to which visual word is, the height of the dimension corresponding to the visual word is increased by 1, until all SIFT descriptor vectors are allocated. In the same way, all the images in the image library are assigned to each dimension of the visual dictionary to obtain the visual word histogram \( H^{(j)}, j = 1, 2, ..., N \) of the pictures in the image library. Among them, \( N \) is the total number of images in the image library.

Step three: Normalize the visual word histogram of all images. Measure the similarity between \( H \) and \( H^{(j)}, j = 1, 2, ..., N \), and calculate the similarity score \( \text{Score}_i = [s_1, s_2, ..., s_N] \) of each image in the gallery according to the similarity.

Step four: The feature vector \( \text{codebook} = \{w_1, w_2, ..., w_m\} \) of the query image is quantified into a certain visual word in the BOW dictionary, and the symbol map of the query image is obtained. Project the visual words in the query image symbol map to the X axis and Y axis to obtain the character strings \( x_{\text{query}} \) and \( y_{\text{query}} \). Rotate the query image to get 4 sets of projection strings \( x_{\text{query}1} \) and \( y_{\text{query}1} \), \( x_{\text{query}2} \) and \( y_{\text{query}2} \), \( x_{\text{query}3} \) and \( y_{\text{query}3} \), \( x_{\text{query}4} \) and \( y_{\text{query}4} \).

Step five: The SIFT feature of the gallery image is quantified as a visual image, and the gallery symbol library is obtained. Project the visual words of each image in the symbol library to the X axis and Y axis to obtain the character strings \( x_{\text{lib}} \) and \( y_{\text{lib}} \), \( j = 1, 2, ..., N \).

Step six: Construct the longest visual word string. Define \( Lcs(str_1, str_2) \) as the longest common substring of strings \( str_1 \) and \( str_2 \). \( \text{patternX}(i) \) and \( \text{patternY}(i) \) are the longest common substrings in the X-axis and Y-axis directions respectively, \( \text{patternMax}(i) \) is the longest common substring in the X-axis direction and the Y-axis direction, \( \text{pattern} \) is the longest string in \( \text{patternMax}(i) \). The specific definition is as follows:

\[
\begin{align*}
\text{patternX}(i) &= Lcs(x_{\text{query}1}, x_{\text{lib}}^i) \\
\text{patternY}(i) &= Lcs(y_{\text{query}1}, y_{\text{lib}}^i) \\
\text{patternMax}(i) &= \max(\text{patternX}(i), \text{patternY}(i)) \\
\text{pattern} &= \max(\{\text{patternMax}(i)\}, i = 1, 2, 3, 4)
\end{align*}
\] (1-4)

Step seven: Calculate the score of each image in the gallery. The score of the gallery image is defined as the similarity \( \text{sim}(I_{\text{query}}, I_{\text{lib}}^j) \) between the query image and the gallery image \( j \),

\[
\text{sim}(I_{\text{query}}, I_{\text{lib}}^j) = \sum_{k=1}^{\text{length(pattern)}}\text{idf}(\text{pattern}(k)), j = 1, 2, ..., N
\] (5)

Among them, \( \text{idf} \) is the weight of each visual word in the bag of words model. According to the similarity score, the similarity score \( \text{Score}_i = [s_1, s_2, ..., s_N] \) of each image in the gallery is obtained.

Step eight: Combining the similarity scores \( \text{Score}_1 = [s_1, s_2, ..., s_N] \) and \( \text{Score}_2 = [s_1, s_2, ..., s_N] \) of each image in the gallery, and the total similarity score according to the formula...
\[ \text{Score} = \lambda \text{Score}_1 + (1-\lambda) \text{Score}_2 \]

\( \lambda \) is an adjustable weight parameter with a value range of 0-1. Sort according to the total score, and finally get the search result.

4. Experimental verification and analysis

Build an image experimental test library firstly. Extract and process 4460 gallery collections in 9 categories, such as geographic atlas, electrical engineering atlas, aviation industry, and mechanical industry as shown in Figure 4.

![Figure 4. Image experiment test](image)

For each type of gallery collection, extract 10 original images of positive samples, and then perform left rotation, right rotation, reduction, enlargement, partial, illumination enhancement, and illumination reduction processing on the original images as shown in Figure 5, resulting in a total of 9 \( \times \) 10 \( \times \) 7,630 positive sample images, and the remaining 4370 images in the test gallery are negative sample images.

![Figure 5. Positive sample generation](image)

The test library used in this experiment has 8 images similar to each positive sample. Therefore, the retrieval accuracy \( TP \) is defined as the proportion of the number of positive example images that are similar to the test image in the first 8 images retrieved in the testing process. In the experiment, the BOW image retrieval model and the BoW optimization model based on the longest common word string constraint were tested respectively. The results are shown in Table 1 and Table 2.

| TP       | Image 1 | Image 2 | Image 3 | …… | Image 9 | Image 10 | Average TP | The total average TP is |
|----------|---------|---------|---------|-----|---------|----------|------------|------------------------|
| Geography| 0.625   | 1       | 0.875   | …… | 0.75    | 0.5      | 0.75       | 0.7625                 |
| Electrician| 0.5     | 0.875   | 0.875   | …… | 0.875   | 0.625    | 0.7625     |                        |

(a) original image          (b) left rotation          (c) right rotation          (d) reduction
(e) enlargement          (f) partial          (g) illumination enhancement          (h) illumination reduction

Figure 5. Positive sample generation
| TP          | Image 1 | Image 2 | Image 3 | …… | Image 9 | Image 10 | Average TP |
|------------|---------|---------|---------|-----|--------|----------|------------|
| Geography  | 0.625   | 1       | 0.875   | …… | 0.875  | 0.75     | 0.775      |
| Electrician| 0.75    | 0.875   | 0.875   | …… | 0.875  | 0.75     | 0.8        |
| Aviation   | 0.625   | 0.75    | 0.875   | 0.75| 0.625  | 0.875    | 0.725      |
| History    | 0.75    | 0.625   | 0.75    | …… | 0.75   | 0.75     | 0.775      |
| Machinery  | 0.625   | 0.875   | 0.75    | 0.75| 0.625  | 0.75     | 0.8        |
| History    | 0.75    | 0.625   | 0.75    | 0.75| 0.625  | 0.875    | 0.825      |
| Medicine   | 0.5     | 0.625   | 0.75    | …… | 1      | 0.75     | 0.675      |
| Horticulture| 0.875  | 0.75    | 0.75   | …… | 0.875  | 0.875    | 0.8625     |
| Plants     | 1       | 0.875   | 0.75   | …… | 0.875  | 0.625    | 0.7625     |

The total average TP is 0.77083

Compared with the BOW model, the accuracy rate of the BoW optimization model based on the longest common word string constraint is increased by 1.834%.

5. Summary
The BOW model has the problem of ignoring the spatial position relationship in the original image. In order to solve this problem, this paper adds the constraint of the longest common visual word string based on the BoW model. The algorithm represents images at the vocabulary level and phrase level at the same time, improves the spatial and semantic relationship between "visual dictionaries". Experiments verify that compared with the BOW model, the retrieval accuracy of the BoW optimization model based on the longest common word string constraint is well optimized.

Acknowledgments
This study was supported by the Major Science and Technology Innovation Project of Shandong (Grant No. 2019JZZY010407).

References
[1] Bagri, N., Johari, PK. (2015) A comparative study on feature extraction using texture and shape for content based image retrieval. International Journal of Advanced Science and Technology, 2: 41-52.
[2] Sivic, J. and Zisserman, A. (2003) Video Google: A text retrieval approach to object matching in videos.
[3] Yu, J., Qin, Z., Wan, T., et al. (2013) Feature integration analysis of bag-of-features model for image retrieval. Neurocomputing, 120:355-364.
[4] Nagaraja, S., Prabhakar, CJ. (2015) Low-level features for retrieval based on extraction of directional binary patterns and its oriented gradients histogram. Computer Science, 2:13-28
[5] Yulei, W., Shusheng, B., Minglei, S., et al. (2014) Image retrieval algorithm based on SIFT, K-Means and LDA. Journal of Beijing University of Aeronautics, 40:1317-1322.
[6] Jegou, H., Douze, M., Schmid, C. et al. (2010) Aggregating local descriptors into a compact image representation, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, pp. 3304-3311.
[7] Zhang, Y., Jia, Z., Chen, T. (2011) Image retrieval with geometry-preserving visual phrases. IEEE Conference on Computer Vision and Pattern Recognition.
[8] Passalis, N. Anastasios, T. (2016) Entropy Optimized Bag-of-Words Representation for Information Retrieval. IEEE Transactions on Knowledge and Data Engineering.

[9] Li, J. Allinson, NM. (2008) A comprehensive review of current local features for computer vision. Neurocomputing, 71:1771-1787.

[10] Wan, S. Lin, P. Yue, L. (2009) An approach for image retrieval based on visual saliency. Image Analysis and Signal Processing, IASP 2009. International Conference on. IEEE, pp.172-175.

[11] Jegou, H. Douze, M. Schmid, C. (2008) Hamming embedding and weak geometric consistency for large scale image search. Computer Vision-ECCV 2008. Spring Berlin Heidelberg, pp.304-317.