LETTER *Special Section on Data Engineering and Information Management*

Image Retrieval with Scale Invariant Visual Phrases

Deying FENG[†a)], ***Student Member***, **Jie YANG**[†], **Cheng YANG**[†], ***and*** **Congxin LIU**[†], ***Nonmembers***

SUMMARY We propose a retrieval method using scale invariant visual phrases (SIVPs). Our method encodes spatial information into the SIVPs which capture translation, rotation and scale invariance, and employs the SIVPs to determine the spatial correspondences between query image and database image. To compute the spatial correspondences efficiently, the SIVPs are introduced into the inverted index, and SIVP verification is investigated to refine the candidate images returned from inverted index. Experimental results demonstrate that our method improves the retrieval accuracy while increasing the retrieval efficiency.

key words: *image retrieval, visual phrases, spatial correspondences, inverted index*

1. Introduction

Content-based image retrieval has attracted increasing interests in recent years. Given a query image, the image retrieval system returns the similar images from an image database. Most state-of-the-art retrieval systems [1]–[3] are based on bag-of-visual-words (BoV) model, in which the images are represented by visual word histograms. Although the BoV model is simple and effective, it does not consider spatial information in the histogram representation. Since spatial information is an important component of image, developing a retrieval method with spatial information is of great significance.

Recently, many algorithms have been proposed to utilize spatial information based on the BoV model. Geometric verification methods [1], [3] calculate spatial information in the post-processing step. Because geometric verification methods are computationally expensive, they are applied only to the top-ranked retrieved images. On the other hand, many approaches encode spatial information in the searching step. Spatial pyramid matching [4] repetitively partitions the image into finer grids, but the rigid spatial information is not invariant to typical transformations. Bundled features method [5] encodes spatial information in the stable regions, but the relative ordering relationship in these regions is sensitive to rotation transformation. Spatial-bag-of-features method [6] arranges the order of visual word histograms to represent the relative spatial relationship, but the arrangement does not correspond to the true transformation. Visual phrase [7] captures spatial information through the co-occurring visual words in the local neighborhood, but sophisticated mining algorithm used in the method discards some representative visual phrases. Geometry-preserving visual phrase (GVP) [8] uses the grids to capture the local and long range spatial layouts of visual words. Due to the limitation of grids, GVP is only invariant to translation transformation. To our best knowledge, most existing retrieval methods encode spatial information after detecting local features. Because the spatial relationship among the detected features is not fixed, few retrieval methods have addressed encoding the spatial information which is invariant to translation, rotation and scale transformation.

In this letter, we propose Scale Invariant Visual Phrases (SIVPs) for content-based image retrieval. A SIVP is a set of visual words which encode the co-occurring features generated by scale invariant feature detection, so it captures the fixed spatial relationship which provides translation, rotation and scale invariance. The SIVPs are then employed to determine the spatial correspondences between query image and database image. To compute the spatial correspondences efficiently, the SIVPs are introduced into the inverted index to count the number of spatial correspondences. Afterwards, SIVP verification is investigated to verify the spatial correspondences and refine the candidate images returned from inverted index. The final retrieved images are the candidate images with the greatest spatial similarity.

The remaining of this letter is organized as follows. Section 2 details our method using scale invariant visual phrases. Section 3 presents experiments performed to evaluate our method. Section 4 summaries this letter.

2. The Proposed Method

2.1 SIVP Generation

The basic idea of SIVP is to generate the co-occurring features in the scale invariant feature detection and encode these features into the co-occurring visual words. The co-occurring features consist of dominant feature and affiliated feature, as shown in Fig. 1. The dominant feature P is detected by scale invariant feature detector and denoted as $P(X, \sigma, \theta)$, where X is the location, σ is the scale, and θ is the orientation. For the first affiliated feature P_1, its location is determined by moving a distance $n\sigma$ along the orientation θ from the location X, its scale and orientation are taken the same as those of dominant feature P. After the first affiliated feature is generated, the fixed spatial relationship is created

Manuscript received June 29, 2012.
Manuscript revised October 12, 2012.

[†]The authors are with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China.
a) E-mail: fdy629@163.com

DOI: 10.1587/transinf.E96.D.1063

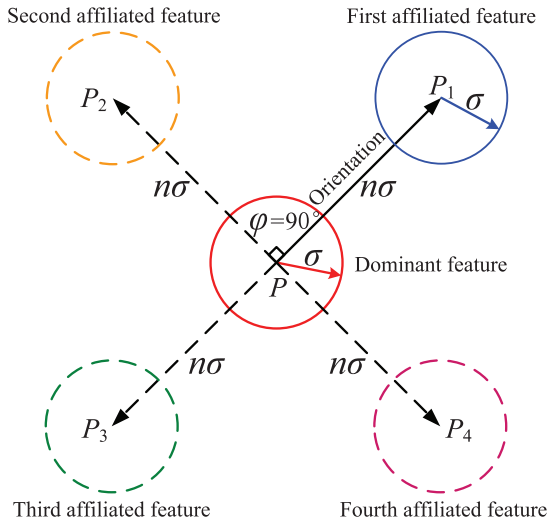

Fig. 1 An example of the co-occurring features generation where the rotation angle φ is set as 90 degree.

between the dominant feature P and the first affiliated feature P_1, and it is invariant to translation, rotation and scale transformation.

Other than the first affiliated feature, more affiliated features can be generated by rotating the orientation of dominant feature. Assuming that the rotation angle is φ, the number of affiliated features is computed as $n_a = 360/\varphi$. Although different rotation angles generate different numbers of affiliated features and change the spatial arrangement of affiliated features, they do not affect the spatial relationship between dominant feature and each affiliated feature. Because only the fixed spatial relationship is used to determine the spatial correspondences, the rotation angle φ can be set as any value. To demonstrate the co-occurring features generation, Fig. 1 gives an example where the rotation angle is set as 90 degree. After the dominant feature and affiliated feature are quantized into the co-occurring visual words, the fixed spatial relationship is integrated into SIVP, which can provide translation, rotation and scale invariance.

To locate the affiliated feature properly, the scale error of dominant feature Δ is introduced into the moving distance $n\sigma$ to analyze the value n. The scale error Δ is caused by scale invariant feature detector, and it is defined as the difference between measured scale σ and true scale σ'. According to the definition, the moving distance $n\sigma$ is computed as:

$$n\sigma = n(\sigma' + \Delta) = n\sigma' + n\Delta \tag{1}$$

As the value n increases, the error $n\Delta$ becomes greater, which decreases the location accuracy of affiliated feature. On the contrary, if the value n is too small, the affiliated feature is close to the dominant feature, and the spatial relationship cannot be described clearly. Thus, the value n is empirically set as 1.5.

After SIVP is generated, it is compared with BoV and GVP and has the following fundamental differences:

First, the three methods capture the spatial relationship in different stages. BoV does not consider the spatial relationship in the retrieval process. Although GVP utilizes the spatial relationship after detecting the local features, it requires additional computation to capture the spatial relationship. Compared with BoV and GVP, SIVP determines the spatial relationship in the process of feature detection and reduces the computational cost.

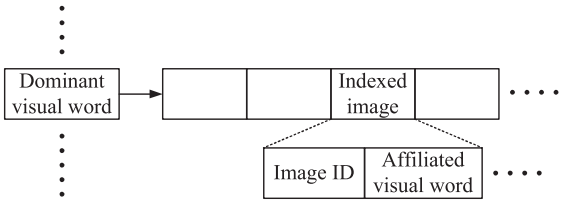

Fig. 2 The structure of inverted index using SIVP.

Second, GVP and SIVP employ different ways to capture the spatial relationship and determine the geometric invariance. GVP uses the grids to capture the spatial relationship among the detected features, but it is only invariant to translation transformation. By contrast, SIVP captures the spatial relationship by generating the co-occurring features in the scale invariant feature detection. Thus, SIVP is invariant to translation, rotation and scale transformation, and it can tolerate more geometric transformations than GVP.

2.2 Inverted Index with SIVP

After generating the SIVPs, the spatial correspondences between query image and database image can be determined if the co-occurring features are represented by the same SIVP. To compute the number of spatial correspondences, the SIVPs are introduced into the inverted index to record the co-occurrences of SIVPs between query image and database image.

Let $D = \{D_t\}_{t=1}^{N}$ be N database images. The database image D_t is represented by bag of SIVPs $V_{D_t} = \{v_i | v_i = (v_i^d, v_i^a)\}_{i=1}^{n_t}$, where v_i is the i-th SIVP, v_i^d and v_i^a are the i-th dominant and affiliated visual word. The inverted relationship between SIVP v_i and database image D_t can be represented as $v_i^d \rightarrow (D_t, v_i^a)$. According to the inverted relationship, the inverted index is constructed by supplying an entry for the dominant visual word v_i^d and storing the database image ID and affiliated visual word v_i^a in the inverted list. If the identical SIVP appears many times in the same image, it is only recorded once in the same inverted list. Figure 2 shows the structure of the inverted index using SIVP.

The query image Q is represented by bag of SIVPs $V_Q = \{v_{q,i} | v_{q,i} = (v_{q,i}^d, v_{q,i}^a)\}_{i=1}^{n_q}$, where $v_{q,i}$ is the i-th SIVP, $v_{q,i}^d$ and $v_{q,i}^a$ are the i-th dominant and affiliated visual word. When the i-th SIVP $v_{q,i}$ is searched in the inverted index, the dominant visual word $v_{q,i}^d$ is used to find the corresponding entry, and the affiliated visual word $v_{q,i}^a$ is used to refine database images in the inverted list. Only the database images which contain the same affiliated visual word are selected from the inverted list, and the occurrences of database

images are accumulated as the number of spatial correspondences. After all the SIVPs of query image are searched in the inverted index, the occurrences of all the database images are ranked and the top-ranked database images are taken as candidate results.

2.3 SIVP Verification

When the inverted index computes the number of spatial correspondences, it sometimes fails to obtain the accurate results, because the orientation of the co-occurring features is not involved in the SIVP. The co-occurring features which contain different orientations may be represented by the same SIVP, which results in false spatial correspondence between query image and database image. To verify the spatial correspondences, SIVP verification estimates the orientation consistency between query image and candidate image. The SIVP verification is summarized as follows:

First, because one pair of co-occurring features in an image may correspond to multiple pairs of co-occurring features in another image, the overall spatial correspondences $C = \{c_j\}_{j=1}^{n_c}$ are divided into single correspondences C_s and multiple correspondences C_m.

Second, a spatial correspondence is randomly selected from single correspondences C_s, and the corresponding orientations θ_q and θ_c are taken as reference orientations for query image and candidate image. To estimate the orientation consistency, the angles between reference orientation and any other orientation are respectively computed in the query image and candidate image. The angles are denoted as $a_q = \{a_{q,j}\}_{j=1}^{n_c-1}$ and $a_c = \{a_{c,j}\}_{j=1}^{n_c-1}$, where $a_{q,j}$ and $a_{c,j}$ are respectively the j-th angle in the query image and candidate image.

Third, based on the spatial correspondences, the error between angle $a_{q,j}$ and angle $a_{c,j}$ is computed as:

$$e_j = |a_{q,j} - a_{c,j}| \tag{2}$$

If the angle error $e_j \leq \eta$, the corresponding orientations are consistent with each other, and the number of positive spatial correspondences n_p is accumulated. The threshold η is empirically set as 5.

Fourth, the process is repeated until a predefined number of iterations are reached. The greatest value $\max(n_p)$ is taken as the spatial similarity between query image and candidate image. After re-ranking candidate images through the spatial similarity, the candidate image with the greatest spatial similarity is the final retrieved result.

3. Experiments and Results

In the experiments, two image datasets are used to test our method. The magazine dataset [9] includes 7665 images scanned from the magazines and 300 query images captured by different camera phones. Because the query images are taken under various conditions, they are affected by many factors such as rotation, scale, illumination and noise contamination. The ImageNet dataset [10] contains 100 K images crawled from the Internet, and it is added as distractors to perform the large scale experiment.

In the image datasets, to ensure the computational efficiency of feature detection, fast Hessian detector is used to detect the dominant features, and SURF descriptor [11] is used to describe the dominant and affiliated features. Because excessive co-occurring features increase the computational cost in the searching process, the co-occurring features are generated only by the first affiliated feature. There are 6 million pairs of co-occurring features extracted from the magazine dataset and 80 million pairs of co-occurring features extracted from the ImageNet dataset. Afterwards, 6 million dominant features are selected from the magazine dataset and clustered by approximate K-means [3] to build the visual word vocabulary. The visual words are utilized to encode the co-occurring features and generate the SIVPs.

To evaluate the retrieval performance, SIVP is compared with BoV [3] and GVP [8]. RANSAC [12] is involved in the BoV and GVP as a post-processing step. All the methods are performed on PC with Intel Dual Quad-Core Xeon E5653 2.53 GHz and 24 GB of RAM. Because the magazine dataset has only one correct database image that corresponds to each query image, we just consider the top one retrieved image. The retrieval accuracy is defined as:

$$Accuracy = D_q / T_q \tag{3}$$

where D_q is the number of queries that obtain positive database images, T_q is the total number of query images.

Figure 3 demonstrates that SIVP provides rotation and scale invariance, and that BoV and GVP cannot achieve the same result. In Fig. 3, three pairs of matching images are respectively generated by BoV, GVP and SIVP. In each pair of matching images, the left image is the query image, and the right image is the corresponding database image. There exists rotation and scale transformation between query image and database image. In Fig. 3 (a), BoV only generates the feature point correspondences and does not consider the spatial relationship, so it cannot provide rotation and scale invariance. In Fig. 3 (b), from the changed spatial relationship and the false spatial correspondences, it can be seen that GVP is not invariant to rotation and scale transformation. In Fig. 3 (c), according to the invariant spatial relationship and the positive spatial correspondences, it can be observed that SIVP provides rotation and scale invariance.

To evaluate the retrieval accuracy of our method, the accuracy is compared under different numbers of visual words in the magazine dataset, as shown in Fig. 4. In Fig. 4 (a), the three methods consider the initial results returned from inverted index. BoV only represents the image as visual word histogram and does not consider the spatial information. Compared with BoV, SIVP captures the spatial relationship and uses the spatial correspondences to compute the image similarity. Thus, SIVP can achieve greater accuracy than BoV. Although SIVP and GVP capture the spatial relationship, SIVP can tolerate more geometric transformations than GVP, thereby determining more positive spatial correspondences than GVP and increasing the im-

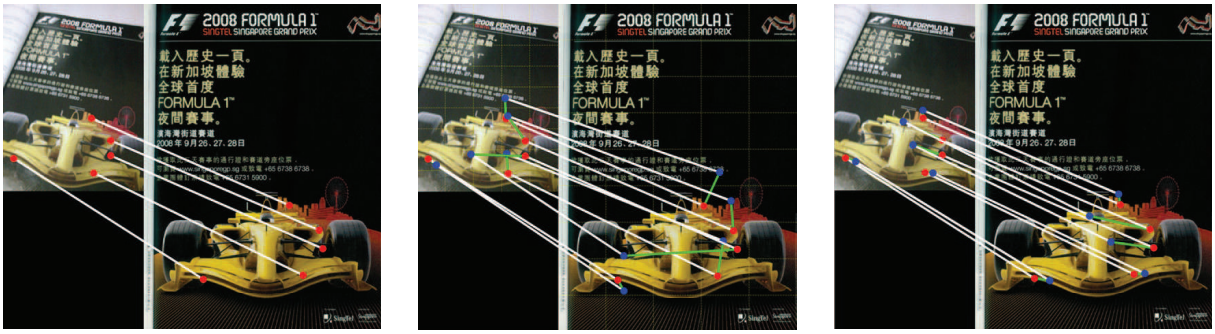

Fig. 3 A set of matching images which are used to compare rotation and scale invariance among three methods: (a) BoV (b) GVP (c) SIVP. In Fig. 3, the red point represents the dominant feature, the blue point represents the affiliated feature, the green line represents the relative spatial relationship between dominant feature and affiliated feature, and the white line represents the corresponding relationship between query image and database image. In order to see clearly, we only demonstrate some of the correspondences in the matching images.

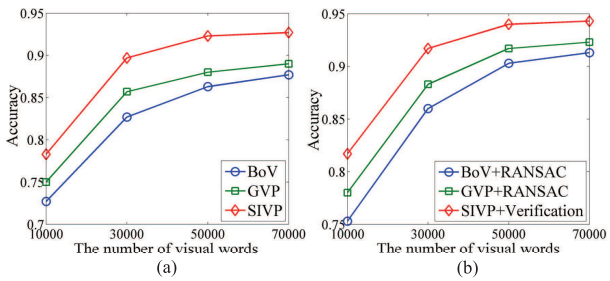

Fig. 4 The accuracy comparison with different numbers of visual words in the magazine dataset: (a) Without the post-processing step (b) With the post-processing step.

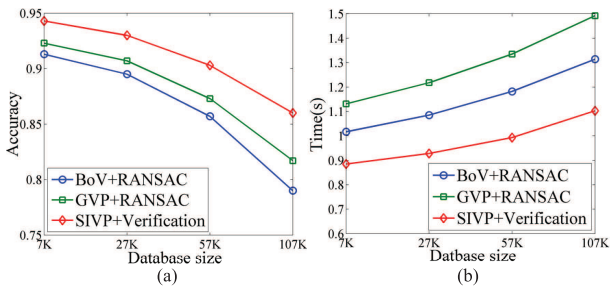

Fig. 5 The accuracy and efficiency comparison with different numbers of distractor images from ImageNet dataset: (a) Accuracy comparison (b) Efficiency comparison.

age similarity. Due to the increased image similarity, SIVP achieves greater accuracy than GVP. In Fig. 4 (b), RANSAC is introduced as post-processing step to estimate spatial correspondences, but it is only applied to the candidate images returned from inverted index. Compared with RANSAC, SIVP is applied to all the database images in the retrieval process. Thus, although the accuracy of BoV and GVP is increased by RANSAC, it is still lower than that of SIVP. From Fig. 4, it can be seen that SIVP achieves greater accuracy than BoV and GVP.

To evaluate the retrieval performance in the large scale image dataset, the accuracy and efficiency are compared under different numbers of distractor images from the ImageNet dataset. In Fig. 5 (a), as the number of distractor images increases, the accuracy of three methods decreases. However, the accuracy of SIVP is still greater than that of BoV and GVP, even if the number of distractor images increases to 100 K. In Fig. 5 (b), as the number of distractor images increases, the time of three methods also increases, but the time of SIVP is much less than that of BoV and GVP. SIVP can improve the retrieval efficiency because it captures the spatial relationship in the feature detection process and only counts the number of spatial correspondences in the inverted index. By contrast, GVP requires additional computation to capture the spatial relationship after detecting the local features, and BoV increases the computational cost because it computes the cosine similarity among the visual word histograms. From Fig. 5, it can be observed that SIVP outperforms BoV and GVP in the large scale dataset.

After comparing the retrieval performance, Fig. 6 shows the retrieved results which are respectively generated by BoV, GVP and SIVP. In Fig. 6 (a), the retrieved images are returned from the magazine dataset. BoV gives the false results in the top-4 retrieved images, because it only computes the cosine similarity among the visual word histograms and does not compare the spatial information of database images. GVP also gives the false results in the top-5 retrieved images, because it cannot tolerate rotation and scale transformation between query image and database image. By contrast, SIVP obtains the positive result in the top-1 retrieved image. Moreover, SIVP gives the false results in the following four retrieved images, because the magazine dataset has only one correct database image for the query image. In Fig. 6 (b), the retrieved images are returned from the magazine and ImageNet dataset. BoV and GVP give the false results in the top-5 retrieved images, because the ImageNet dataset which is added as distractors affects the retrieved results. However, SIVP can still obtain the positive result in the top-1 retrieved image.

Fig. 6 Top-5 retrieved results generated by BoV, GVP and SIVP, and top-1 retrieved results marked by green rectangle: (a) From the magazine dataset (b) From the magazine and ImageNet dataset.

4. Conclusion

We have presented a retrieval method using scale invariant visual phrases. The first advantage is that SIVP provides translation, rotation and scale invariance and employs the spatial correspondences to retrieve images. Secondly, SIVP is integrated into the inverted index, which computes the number of spatial correspondences efficiently. Thirdly, SIVP verification refines the candidate images returned from inverted index and further improves the retrieval accuracy. The experimental results demonstrate that our method outperforms the state-of-the-art methods in the content-based image retrieval.

Acknowledgments

The authors thank the anonymous reviewers for their valuable comments. This work was supported by Committee of Science and Technology, Shanghai (No.11530700200) and National Natural Science Foundation, China (No.61273258).

References

[1] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," IEEE International Conference on Computer Vision, pp.1470–1477, Nice, France, 2003.
[2] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.2161–2168, New York, USA, 2006.
[3] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.1–8, Minnesota, USA, 2007.
[4] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bag of features: spatial pyramid matching for recognizing natural scene categories," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.2169–2178, New York, USA, 2006.
[5] Z. Wu, Q.F. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.25–32, Florida, USA, 2009.
[6] Y. Cao, C. Wang, Z. Li, L.Q. Zhang, and L. Zhang, "Spatial-bag-of-features," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.3352–3359, San Francisco, USA, 2010.
[7] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation pattern: from visual words to visual phrases," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.1–8, Minnesota, USA, 2007.
[8] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.809–816, Colorado Springs, USA, 2011.
[9] D. Feng, J. Yang, and C. Yang, "Efficient indexing for mobile image retrieval," IEEE International Conference on Data Mining Workshops, pp.793–798, Vancouver, Canada, 2011.
[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.248–255, Florida, USA, 2009.
[11] H. Bay, A. Ess, T. Tuytelaars, and L.V. Gool, "Speeded-up robust features (SURF)," Comput. Vis. Image Understand., vol.110, no.3, pp.346–359, June 2008.
[12] M. Fischler and R. Bollers, "Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography," Commun. ACM, vol.24, no.6, pp.381–395, 1981.