Invariant Bag of Words for Image Retrieval

Qi Zhang Frank Anemaet
Leiden University
Leiden Institute for Advanced Computer Science
contact: qifuzhang@gmail.com

ABSTRACT

The avalanche of images from personal devices and steaming servers in the recent years has made the visual search problem timely and important. In this paper, we propose an algorithm which encodes spatial information into the BoV representation. Unlike previous approaches in spatial bag of words, our algorithm forms spatial visual phrases which are invariant to translation and rotation. Furthermore, we demonstrate that our algorithm outperforms other recently proposed spatial image retrieval algorithms in terms of accuracy.

Keywords

algorithms, content based image retrieval, bag of visual words, computer vision

1. INTRODUCTION

An explosion of visual media from a multitude of sources has made the search problem timely and important. Due in no small part to the smartphone wave, we have seen images pop up not just in personal collections but over the landscape of social media from Twitter to Facebook and Tumblr. Because a smartphone is also a communication device, it is frequently the only device which is always with the user and thus gives the user the immediate possibility of snapping a picture or recording a video. With the latest generation of everywhere Internet, users can also post or email their pictures to their friends and social communities instantly. Applications of using the visual media are also expanding to new areas such as question and answer systems [18] and emotion recognition [13]. The avalanche of imagery brings with it the question of how to search for it, make sense of the needle in the proverbial haystack.

The progression of content based image retrieval systems [1, 16, 17, 18] created in the last decade has evolved from simple feature comparison to advanced intermediate representations such as the bag-of-visual-words (BoV) model. Despite its popularity the BoV model discards all spatial information that is available in the images. By including spatial information it may be possible to improve image retrieval accuracy.

2. RELATED APPROACHES

The fundamental representation used in many state-of-the-art approaches is to create a Bag of Visual Words (BoV). The general steps are as follows:

(1) Detect key points and extracting *local descriptors*. This step is mandatory since these key points are used to form visual words. Well known competitive methods include SIFT [8], PCA-SIFT [9], SURF [10], ORB [11] and BRISK [12].

(2) Clustering *local descriptors* where each cluster becomes a visual word (for a good introduction see Philbin, et al. [6] and Yang, et. al. [7]). The collection of visual words are stored in a codebook. In a straightforward implementation, one would represent an image by the frequency of all possible visual words. From Philbin, et al. [6], in their experiments the optimal codebook size was 1 million.

(3) Indexing and search. One may apply algorithms from text-retrieval systems, such as the inverted file and also utilize algorithms which perform fast approximate nearest neighbor matching [19].

Recent works propose extensions to the Bag of Visual words model to improve the accuracy and/or performance of CBIR systems. In the work by Zhuang, et al.[3], they model the joint distribution so that a low factor subspace can be derived. This new subspace only requires one-tenth of the storage and gives a competitive performance on a medical testset.

In the work by Grzeszick, et al. [5], they combine local features with spatial information in the clustering process. Their method is similar to the top level of spatial pyramids but computes a special vocabulary which is specific to each region. A comprehensive review is out of the scope of this article. One notable example is

Next, we introduce two competitive algorithms for taking into account spatial information, SPM [2] and GVP [3], which we will use in our performance evaluation.

2.1 Spatial Pyramids

The BoV model introduced above discards any spatial information. Lazebnik, et al. [2] introduced Spatial-Pyramid Matching (SPM) which encodes spatial information. The method had initially been created for recognizing scenes such as highway, office, street, forest etc. The SPM model is inspired by the intuition that people can recognize scenes while overlooking various details - and thus perceive scenes in a holistic way. Thus scenes may be recognized or classified based on the spatial layout of the image while neglecting the details.

An essential concept in SPM is the increasingly finer grids over the feature space. For each level we construct a histogram H by concatenating the cells. Finally, we concatenate the histograms of all levels into one (large) histogram.

SPM is exceptionally well known in the research literature and frequently is used as a benchmark by the spatial bag of words community, which is why it will be used in this study.

2.2 Geometry Preserving Visual Phrases

Zhang, et al [3] proposed an algorithm that extends the Bag-of-visual-words model with spatial information, known as Geometric-preserving Visual Phrases (GVP). In their work they found that the GVP algorithm has a higher precision than the BoV model (even with using RANSAC). The GVP algorithm captures both local and long-range spatial relationships.

The GVP algorithm calculates the offset of the word j in the offset space based on the location in image I and I'. The x value is the I'x subtracted by Ix and likewise the y value is I'y subtracted by Iy. For example, if we wish to find the location of word C in the offset space, we find its location in image I: (3,3) and in I': (5,3). Then we calculate the offset by subtracting I' and I and find (2,0), A visual word may appear both on the positive and negative side of the axis.

3. ROBUST INVARIANT PHRASES

As mentioned in their paper, the GVP algorithm as presented by Zhang et al. [3] is limited to translation invariance. Here we introduce an approach to address rotation invariance.

We define a visual phrase as k visual words in a certain spatial layout. First we create a different offset space than in GVP where only distances are retained. In Figure 1, we show 3 words in a spatial arrangement on images I and I'. We calculate the x and y distances from A to B and A to C which for this example we will say are 2 and 3 respectively. In the offset matrix, we then have an entry for ABC at position (2,3). Because we are only keeping the distance information, this allows rotation within an area in their image space.

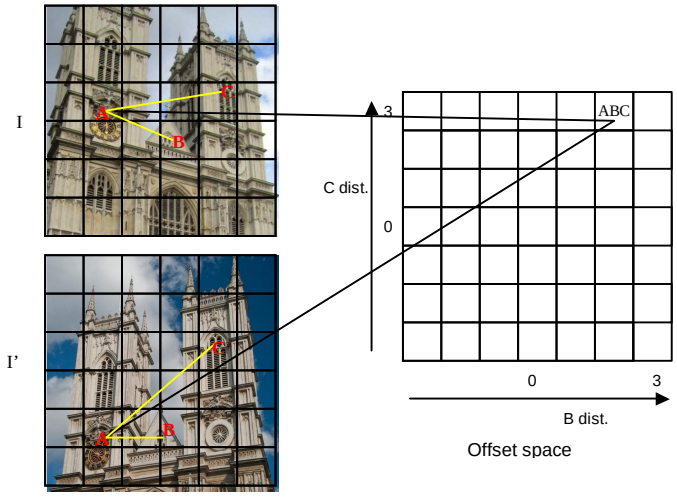

Figure 1. An illustration of creating an offset space for RIP.

Thus, a if we have a visual phrase of length k=2 it will be found in this model - independent of where the second word is located as long as it is within the same distance.

In order to limit the search space, we perform analysis on a large set of words to determine how informative they are. In the literature, the *tf-idf* (term frequency–inverse document frequency) is often used to determine which words are more interesting or informative. We expand upon tf-idf to also examine pairs and triplets of words. This gives us three ranked lists, R, which corresponds to the most interesting single words, pairs of words, and triplets of words.

For each image in the database, we

(1) For the single words list, we only used the typical tf-idf weighting.

(2) Create an MxM offset space similar to Figure 1 based on R for pairs, and triplets: 1-dimensional for pairs; 2-dimensional for triplets.

For the query image,

(3) We also do steps (1)-(2) above and compare the offset matrices from the query image to the offset matrices of the database image. Each shared instance in the offset matrices counts as a single vote. Each shared entry in the single words list counts as a single vote.

(4) Similarity is measured by the number of votes.

4. EXPERIMENTS

In these experiments we used two datasets: WWW-40K (created by scraping 40,000 images from the WWW using Google image search) and the well known NUS-WIDE (269,648 images) dataset with the SIFT salient points. Significant preprocessing had to be performed. Specifically, computing the clusters and pairs and triplets ranking lists was computationally expensive. To keep the memory costs tractable, we computed the single words ranked list first, then only considered a subset of the top 300,000 according to tf-idf for the pairs ranked list. Using the 300,000 top elements from the pairs and singles lists, we then computed the triplets. We had shared access to an Intel Blade Server with 24 2Ghz processors and 192GB RAM for the clustering (~4 days) and computation of an approximation to the pairs and triplets ranking lists (~11 days). The noted times are approximate and are upper-bounds because there were other users on the server. For the offset matrix, M=7. Examples of test images are in Figure 2.

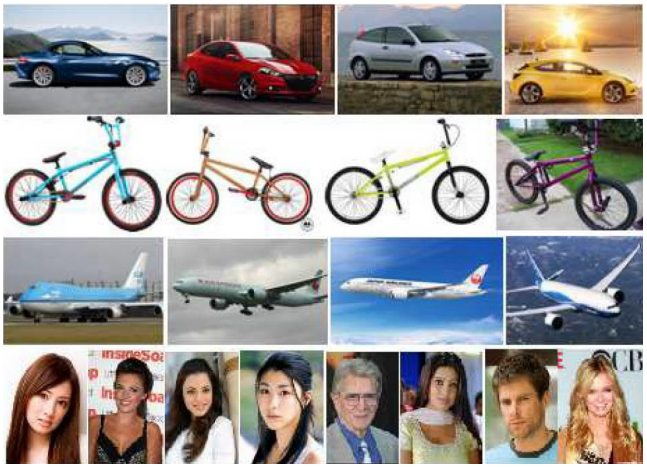

Figure 2. Examples images from top row to bottom for the classes of car, bicycle, plane, and person .

In the first test, we measured the mean average precision on the WWW-40K which consists of 1000 images in 10 categories with 1000 relevant images per category and 30,000 other images as noise. The codebook size of the BoV was 1 million (based on Philbin, et al. [6]) and was computed using the approximate k-means algorithm. The results were averaged over using each image in a category as a query and are shown in Figure 3.

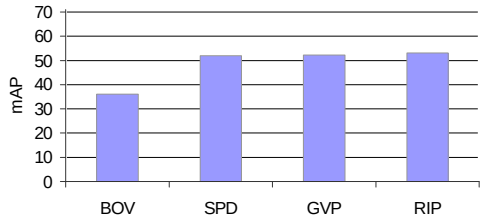

Figure 3. Mean average precision over all of the methods for the WWW-40K.

In the second test, we used 25,000 images from the fully annotated MIRFLICKR [14] with the NUS-WIDE [15] as distractor images. We used the annotations given by the testset authors for calculating mean average precision for 40 categories (e.g. sky, night, people, plant, structure, indoor, clouds, food, tree, etc.).

On a Core i5 3Ghz computer, results were shown to a new query in less than 1 second. This is acceptable performance but it was 73% slower than BoV.

In measuring the performance of the different approaches, we used the mean average precision (mAP). Figure 4 displays a subset of the results for RIP for popular categories.

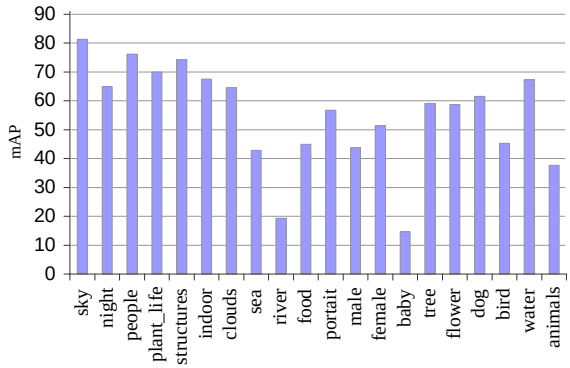

Figure 4. Mean average precision using RIP by category

We also compared the total mAP for the four approaches and the results are shown in Figure 5. The generic bag of visual words (BoV) algorithm had the worst accuracy on all data sets in terms of precision. The Spatial Pyramid (SPD) algorithm significantly improved the accuracy in general as compared to BoV. Both GVP and RIP outperformed SPM in terms of the mAP for both tests. The RIP algorithm had the best overall performance for our experiments.

The RIP algorithm has its share of weaknesses. First, it was significantly slower than the classic BoV. Second, it requires significantly more preprocessing of the database beforehand than the other methods. 192GB RAM is not a typical amount for a normal desktop or workstation. On the other hand, some of our high performance colleagues have indicated that there are several possibilities for sparse matrix computing which could make it computable on a 32GB desktop PC.

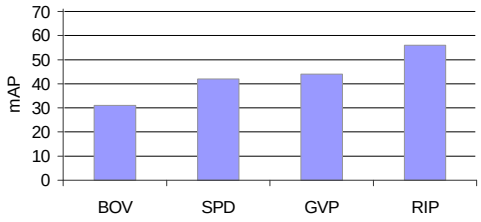

Figure 5. Mean average precision over all of the methods for the MIRFLICKR/NUS-WIDE.

Even if we succeeded, the preprocessing requirements could render RIP to be unsuitable for applications where images need to be searched immediately or where there is a steady stream of new imagery such as in social media clouds like Twitter.

In Figure 6 we presented a screenshot of the results of using RIP for several different queries. Another detail to note is that the images were converted to grayscale for direct processing by SIFT (which does not use color). Thus, even though color images are shown in the Figure, RIP and the other competitive methods were only given the grayscale versions.

One can observe the kinds of false positives which RIP returns. Cars, planes and portraits were likely confused by RIP due to those pictures frequently having a simple background and salient points linked to specific local curved shapes.

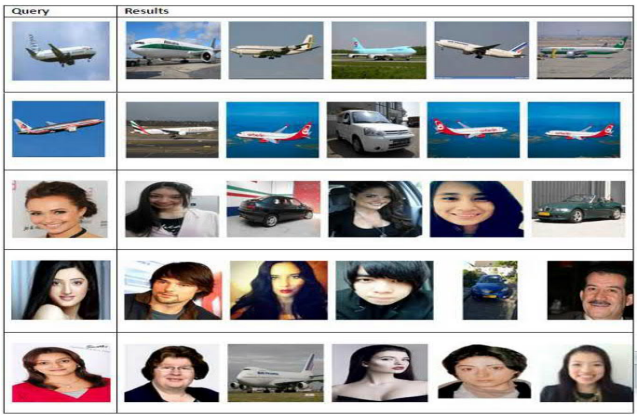

Figure 6. Results in using the RIP algorithm.

5. DISCUSSION AND CONCLUSIONS

Our contribution is an algorithm called Robust Invariant visual Phrases (RIP) for bag of visualwords (BoV) based image retrieval. RIP is meant to address spatial constraints in BoV methods. Our approach was based on tf-idf weighted word pairs and triplets and distance constrained patterns. This gave our approach tolerance for translation and rotation transformations. The BoV had the lowest precision on average for the datasets. The GVP algorithm consistently outperformed BoV and SPD in our tests. We also found that SPD had a shorter computation time than GVP and RIP. In our experiments, RIP outperformed BoV, SPM and GVP in terms of mean average precision.

Why did RIP perform better than GVP? There are several significant differences between RIP and GVP. Even though both use offset matrices, RIP uses the distance, not the x and y offsets. This allows RIP to be more tolerant of rotation changes than GVP such as in the example in Figure 1. Furthermore, RIP combines the results of several offset matrices specifically the pairs and triplets which might contribute to a more stable solution. Third, the pre-computation of the pairs and triplets ranked lists allow RIP to focus on the most informative word patterns. Fourth, the test sets should not be neglected. It is clear that the GVP was designed to be invariant only to translation. So, for datasets which contain only translation within the images in a category, we suspect that GVP will outperform RIP.

Because we had several approaches and two datasets at hand, we were also curious about the brittleness of the codebooks. Specifically, what happens if we use the NUS-WIDE codebook for the WWW-40K test? Is NUS-WIDE large enough to create a representative codebook for generic WWW image datasets?

In Figure 7, we show the results of using the NUS-WIDE codebook for the WWW-40K.

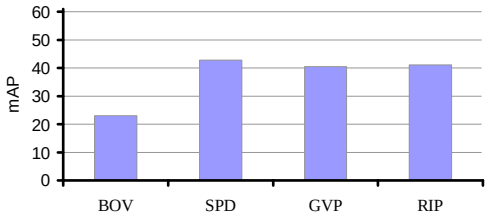

Figure 7. mAP when using NUS-WIDE codebook for WWW-40K

For SPD, GVP and RIP the mAP decreases by an average of 13.4. Using the NUS-WIDE codebook also results in RIP having a slightly lower mAP than SPD. This means that it may be challenging to create a universal codebook which is something to investigate for the future.

In the future we will also be determining if the pre-computation of the ranked lists can be done on a Desktop PC. Furthermore, we think it is important to consider the effect of the salient point detector. Although we used the well known SIFT, it is possible that either a newer detector such as BRISK might give substantial performance benefits or designing a salient point detector specifically for pairs and triplets might be fruitful.

6. ACKNOWLEDGMENTS

Our thanks to Leiden University. This work was based on GVP [3] and the work by Frank Anemaet in his masters research as starting points.

7. REFERENCES

[1] R. Datta, D. Joshi, J. Li, J. Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (CSUR), Volume 40 Issue 2.

[2] S. Lazebnik, C. Schmid, J. Ponce. 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. IEEE Computer Vision and Pattern Recognition, pp. 2169-2178.

[3] Yimeng Zhang, Zhaoyin Jia, Tsuhan Chen. 2011. Image retrieval with geometry-preserving visual phrases. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 809-816.

[4] X. Zhuang, S. Wu, P. Natarajan. 2013. Compact bag-of-words visual representation for effective linear classification. ACM International Conference on Multimedia, pp. 521-524.

[5] R. Grzeszick, L. Rothacker, G. Fink. 2013. Bag-of-Features Representations Using Spatial Visual Vocabularies for Object Classification. IEEE International Conference on Image Processing (ICIP), Melbourne, Australia.

[6] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8.

[7] Jun Yang, Yu G. Jiang, Alexander G. Hauptmann, Chong W. Ngo. 2007. Evaluating bag-of-visual-words representations in scene classification. Proceedings of the ACM International Workshop on Multimedia Information Retrieval (MIR), pp. 197-206.

[8] David G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, Vol. 60, No. 2. (1 November 2004), pp. 91-110.

[9] Yan Ke, R. Sukthankar. 2004. PCA-SIFT: a more distinctive representation for local image descriptors. IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2 (2004), pp. 506-513.

[10] Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool. 2008. Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3. (June 2008), pp. 346-359.

[11] E. Rublee, V. Rabaud, K. Konolige, G. Bradski. 2011. ORB: An effcient alternative to SIFT or SURF. International Conference on Computer Vision (ICCV), pp. 2564-2571.

[12] S. Leutenegger, M. Chli, and R. Siegwart. 2011. BRISK: Binary Robust Invariant Scalable Keypoints. International Conference on Computer Vision (ICCV), pp. 2548-2555.

[13] Y. Sun, N. Sebe, M. S. Lew and T. Gevers. 2004. Authentic Emotion Detection in Real-Time Video. International Workshop on Human-Computer Interaction, Lecture Notes in Computer Science, vol. 3058, Springer (2004), pp. 92–101.

[14] M. J. Huiskes, B. Thomee, M. S. Lew. 2010. New Trends and Ideas in Visual Concept Detection. ACM International Conference on Multimedia Information Retrieval (MIR), Philadelphia, USA.

[15] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. 2009. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. ACM International Conference on Image and Video Retrieval (CIVR). Greece. Jul. 8-10.

[16] M. S. Lew, N. Sebe, J. P. Eakins. 2002. Challenges of image and video retrieval. Conference on Image and Video Retrieval (CIVR). London. pp. 1-6.

[17] N. Sebe, M. S. Lew, A. W. M. Smeulders. 2003. Video Retrieval and Summarization. Computer Vision and Image Understanding, pp. 141-146.

[18] B. Thomee, M. S. Lew. 2012. Interactive Search in Image Retrieval: A Survey. International Journal of Multimedia Information Retrieval, Springer, 1(2), pp. 71-86.

[19] S. O'Hara, B. Draper. 2013. Are you using the right approximate nearest neighbor algorithm. IEEE Workshop on Applications of Computer Vision (WACV), pp. 9-14