Spatial-Content Image Search in Complex Scenes

Jin Ma Shanmin Pang✉
Xi'an Jiaotong University
m799133891@stu.xjtu.edu.cn pangsm@xjtu.edu.cn

Bo Yang
Xi'an Polytechnic University
yangboo@stu.xjtu.edu.cn

Jihua Zhu Yaochen Li
Xi'an Jiaotong University
{zhujh, yaochenli}@xjtu.edu.cn

Abstract

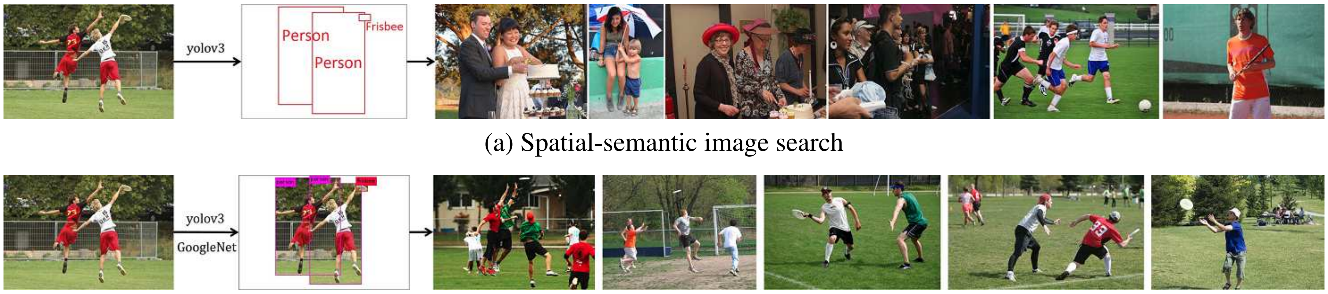
Although the topic of image search has been heavily studied in the last two decades, many works have focused on either instance-level retrieval or semantic-level retrieval. In this work, we develop a novel visually similar spatial-semantic method, namely spatial-content image search, to search images that not only share the same spatial-semantics but also enjoy visual consistency as the query image in complex scenes. We achieve the goal by capturing spatial-semantic concepts as well as the visual representation of each concept contained in an image. Specifically, we first generate a set of bounding boxes and their category labels representing spatial-semantic constraints with YOLOV3, and then obtain visual content of each bounding box with deep features extracted from a convolutional neural network. After that, we customize a similarity computation method that evaluates the relevance between dataset images and input queries according to the developed image representations. Experimental results on two large-scale benchmark retrieval datasets with images consisting of multiple objects demonstrate that our method provides an effective way to query image databases. Our code is available at https://github.com/MaJinWakeUp/spatial-content.

1. Introduction

Image retrieval has long been an active research topic in computer vision and multimedia as it is essential in many applications, such as online shopping [15], person identification [9] and photo management [23]. The objective of image retrieval is to return a ranked list of images that are relevant to a query within a very large database. Basing on the definition of relevance, we can roughly split literature into two main groups: instance- and category-level image retrieval. In instance-level retrieval, we want to search images of the exact same object presented in a query image. This direction is first tackled with sparse Bag-of-Words (BoW) representation model [25], and then with dense yet compact representation models, such as Fisher Vectors [19] and VLAD [11]. In short, these models represent each image by embedding and aggregating a set of local descriptors (e.g., SIFT) into a global image-level vector. Recently, this direction has benefited from the success of deep learning, and builds image representations[1, 31, 28, 18] either with activations of a fully connected layer or a convolutional layer from a Convolutional Neural Network (CNN). Previous work has achieved excellent results on instance retrieval within a single category, such as buildings [20] and shoes [27]. However, it is infeasible to collect labelled data to train a retrieval model for each class separately as it has been estimated that humans can distinguish at least 30,000 object classes [3].

Category-level image retrieval, which concentrates on searching semantic-related images of the same category[29, 2], is on the other end of the retrieval spectrum. Typical research on category-level retrieval often uses machine learning techniques to learn a mapping between visual and semantic representations, such as image caption generation[8, 17] and word vector embeddings [7, 4]. Despite popularity, such methods cannot deal with spatial constraints such as object positions. Although the retrieval method [30] based on concept maps can deal with spatially distributed semantic concepts, it cannot reflect the relative scales of different objects. Recently, Mai *et al.* [16] tackle the scale problem by manipulating the sizes of concept text-boxes on a 2D query canvas. However, similar to other category-level retrieval methods, it still suffers from the limitation that the retrieved images may not come up to expectation of query users due to visual semantic discrepancy.

In this work, we extend the state-of-the-art category-level retrieval methods, and introduce a novel image retrieval task in complex scenes, where the goal is to search



(a) Spatial-semantic image search

(b) Spatial-content image search

Figure 1. Spatial-content image retrieval. (a) Searching with a set of bounding boxes (generated by YOLOV3 [22]) representing spatial-semantic constraints of the query image returns confused retrieval results due to lack of query visual information. (b) By incorporating visual features into retrieval process, our method return visually similar semantic retrieval results as the query image.

images that not only share the same spatial-semantics but also enjoy visual consistency as the query image. As shown in Fig.1, compared with conventional spatial-semantic image search, our retrieval system can indeed return images that are both semantically and visually relevant to the user query image.

The main challenge in developing such an image search technology is to design appropriate image representations. When humans look at an image containing multiple objects, they usually observe the objects presented in the image, and then observe the position of each object as well as position relations between different objects. Inspired by this observation characteristic, we consider recent advances of object detection techniques and generate a set of bounding boxes representing spatial-semantic constraints with YOLOV3 [22]. Then, we obtain visual content of each bounding box with deep convolutional features extracted from a pre-trained Convolutional Neural Network (CNN). Finally, we represent images with a set of tuples, where each tuple contains information of position, label and visual description of an object. After that, we design a similarity computation method that takes into consideration of both semantic and visual similarities to rank database images. To our knowledge, this is the first work to propose a visually similar based spatial-semantic image retrieval system.

We organize the rest of this paper as follows. Section 2 describes the proposed image retrieval system, where we present the computation of image representations and image relevant scores in detail. We report experimental results as well as some discussions on the MS-COCO and Visual Genome datasets in Section 3. Finally, we conclude the paper in Section 4.

2. Proposed Approach

2.1. Image Representation

Our representation framework consists of two components: spatial-semantic object representation and visual object representation. First, we tackle spatial-semantic object representation with object detection. Given an image, the goal of object detection is to return the spatial location and extent of each object instance usually via a bounding box. The main object algorithms in the literature can be broadly classified into two groups: one-stage and two-stage. The one-stage detection framework refers broadly to architectures that directly predict class probabilities and bounding box offsets from full images with a single feed forward CNN network. Representative algorithms include OverFeat [24], SSD [14] and YOLO [21]. The two-stage detection framework first includes a pre-processing step for region proposal generation, and then determines the category labels of the proposals with category-specific classifiers. Well-known methods include RCNN [6], Fast RCNN [5] and Mask RCNN [10]. In this work, we prefer to one-stage framework, and select YOLOv3 [22] for our purpose as it is both effective and efficient in practice. For a given image M, we choose bounding boxes corresponding to score values greater than 0.5 for final detection. The obtained bounding boxes and category-labels are respectively considered as spatial constraints and semantic information of object instances contained in the image. Therefore, they formulate a spatial-semantic representation of the image, and can be used to support image retrieval. However, as shown in Fig.1 (a), searching only with spatial-semantic information is not enough to return expected similar images. This is because even if the object instances belong to a same category, their visual differences may be still huge. Therefore, it is necessary to leverage visual features to return relevant images what query users expect.

Second, we generate the visual representation of each object instance with deep convolutional features. Specifically, we feed the image M through a convolutional neural network, and correspondingly obtain a 3D tensor vector $\mathcal{F} \in \mathbb{R}^{W \times H \times K}$ produced by the activations (responses) of a convolutional layer, where $W \times H$ is the spatial resolution of the feature maps, and K is the number of feature maps (channels). Basing on the fact that the convolutional layer still preserves the spatial information of the input image, for an object instance detected by YOLOv3 in the original image M, we derive its subset of locations on the spatial grid $W \times H$ using the size and location of the bounding box. Consequently, we obtain the feature set $\mathcal{F}' \in \mathbb{R}^{W' \times H' \times K}$ of the object instance, where $W' \times H'$ is the object size in the feature space. Finally, we compute visual representation $f = (\phi_1, \ldots, \phi_K)^T$ of the object with the simplest sum-pooling strategy. That is,

$$\phi_k = \sum_{p=1}^{W'} \sum_{q=1}^{H'} \mathcal{F}'_{pqk}, \quad \forall k = 1, \ldots, K. \tag{1}$$

Up to now, we have built an effective representation $I^M = \{O_1, O_2, \ldots, O_n\}$ for the image M, where each O_i is an object instance of the image. Besides, each O_i contains three components $\{b_i, l_i, f_i\}$, where b_i, l_i and f_i denote its bounding box, category label and visual vector, respectively. Formally, the image representation of the image M is

$$I^M = \{O_1, ..., O_i, ..., O_n\}, \quad O_i = \{b_i, l_i, f_i\}. \tag{2}$$

2.2. Similarity Computation

Now, we consider how to compute similarity scores between different images with Eq.(2). In [16], the authors define a method to compute the relevance between an input query Q and a retrieved database image D as

$$S(I^Q, I^D) = \frac{1}{|I^Q|} \sum_{O_i \in I^Q} \max_{O_j \in I^D} \mathbb{I}(l_i = l_j) \frac{b_i \cap b_j}{b_i \cup b_j} \tag{3}$$

where $|I^Q|$ denotes the number of object instances in Q; O_i and O_j correspondingly denote the object instances of the image Q and the image D; while l_i and b_i (l_j and b_j) are the category label and bounding box of O_i (O_j), respectively. Besides, $\mathbb{I}$ represents the indicator function which takes the value 1 if its argument is true and zero otherwise. The main idea behind the Eq.(3) is that, for each object instance O_i in the query image Q, it first searches all the object instances with the same category label as O_i in the database image D, and then selects the object instance with the largest spatial overlap ratio to compute the object relevance score. Finally, it obtains the relevance score of Q and D by averaging all the individual object relevance scores.

As stated above, the category-label information required by the relevance score Eq.(3) is too coarse to describe the query user intent, and therefore searching with Eq.(3) cannot guarantee promising retrieval results. To overcome this limitation, we incorporate visual object representation into relevance score computation:

$$\widetilde{S}(I^Q, I^D) = \frac{1}{|I^Q|} \sum_{O_i \in I^Q} \max_{O_j \in I^D} \mathbb{I}(l_i = l_j)(\alpha \frac{b_i \cap b_j}{b_i \cup b_j} + (1 - \alpha) \cos(f_i, f_j)) \tag{4}$$

where $\alpha \in [0, 1]$ is a scale parameter ; f_i and f_j denote the visual representations of object-instances O_i and O_j, respectively; the symbol cos means the *Cosine* similarity of two vectors. We call the score computed by Eq.(4) *spatial-content relevance score* as it adds visual content similarity when matching objects of the query image Q and the database image D.

The parameter α in Eq.(4) is used to balance the contribution of semantic representation and visual representation in similarity computation. When $\alpha = 1$, Eq.(4) degenerates to Eq.(3); while on the other end $\alpha = 0$, the equation becomes

$$\widetilde{S}_{\alpha=0}(I^Q, I^D) = \frac{1}{|I^Q|} \sum_{O_i \in I^Q} \max_{O_j \in I^D} \mathbb{I}(l_i = l_j) \cos(f_i, f_j) \tag{5}$$

which means we compute the relevance score only without the spatial constraint. In other words, we compute the relevance score between Q and D by matching object descriptors belonging the same category labels. We empirically select the optimal value for α according to experimental evaluations.

It should be noted that, although the spatial-semantic constraint is included in both Eqs.(3) and (4), it is obtained in different ways and it serves different aims in these two equations. First, the bounding box and category-label information used in Eq.(3) is obtained with image annotations, while in our Eq.(4), these information is produced by the detection technique YOLOv3. Second, the spatial-semantic constraint is aimed to annotate ground-truth spatial-semantic relevance images in Eq.(3), however it is partly used for searching relevance images in Eq.(4), and its effect will be examined by our experimental results.

As is just said, the relevance score of Eq.(4) is based on the spatial-semantic information produced by the object detection technique. However, it is worth noting that, the number of categories used to train the detection model is limited. This means we may not detect every object for some images. In order to handle this special case, we consider an image as a single object, and use the original deep convolutional features $\mathcal{F}$ to generate its representation with Eq.(1). That is $I = \{O\} = \{f\}$. If this case happens to at least one of two compared images, we compute their similarity score using the following equation:

$$\widehat{S}(I^Q, I^D) = \frac{1}{|I^Q| + \beta} \begin{cases} \max\limits_{O_j \in I^D} \cos(f^Q, f_j), & I^Q = f^Q \\ \max\limits_{O_i \in I^Q} \cos(f^D, f_i), & I^D = f^D \end{cases} \tag{6}$$

where $\beta \in \mathbb{N}$ is a penalization parameter to tune the relevance score between Q and D in this situation. Both the Eqs.(4) and (6) together constitute the similarity computation method designed for the proposed spatial-content image retrieval.

2.3. Flowchart of Spatial-content Image Search

To be clear, we summarize the framework of the proposed spatial-content image retrieval method as follows:

1. For each image D in a large-scale database, we represent it with $I^D = \{O_1, ..., O_i, ..., O_n\}$, where $O_i = \{b_i, l_i, f_i\}$ is the detected object instance in D with YOLOv3. The components b_i, l_i and f_i represent the bounding box, category label and visual representation of O_i, respectively.

2. For a user given query image Q, we generate its representation I^Q, and then compute its relevance score with each image D in the database. Specifically, if both Q and D have object instance labels, we compute their similarity score with Eq.(4). Otherwise, we compute their similarity score with Eq.(6).

3. Return a ranked list of the retrieved images by sorting the relevance scores.

3. Experiments

In this section, we utilize the GNet-Conv feature extracted from the fifth inception convolutional layer of GoogleNet [26] to construct the object visual representation in Eq.(2). For each object, we apply L2 normalization to the feature before and after sum-pooling operation. The GNet-Conv baseline, which uses the L2 normalized sum-pooling vector of whole GNet-Conv feature for image search, is compared to show the effectiveness of our method.

3.1. Standard Relevance Score

In order to evaluate the performance of the proposed method, we need to define a standard relevance score. In [16], Mai *et al.* use Eq.(3) as standard relevance score but without any proof. While in [8], Gordo *et al.* discuss the selection of standard relevance score in detail. Specifically, this publication investigate the agreement score between users' ranking and several visual baseline methods based on dataset annotations, and discover that the tf-idf BoW representation of captioned texts has the highest agreement score with users, which means this representation is a good predictor of the relevance between two images. So in this paper, we employ the cosine similarity of L2 normalized tf-idf BoW representations of two images as their proxy standard relevance score.

3.2. Datasets

We evaluate the performance on two large-scale datasets that are designed for cognitive scene understanding tasks: MS-COCO [13] and Visual Genome [12].

MS-COCO: We use the training and validation sets with captioned texts of MS-COCO 2017 in this experiment. The validation set contains 5000 images, it can be treated as a small dataset for experimental setup. We select 15 groups images of complex scenes from the validation set, every image within one group is similar to each other, then choose 5 images from each group randomly and combine them together into a query set of 75 images. Besides, we filter some images that are similar with these query images from training set according to standard relevance score (if one image's tf-idf similarity with any query is larger than 0.3, then filter it), then add remaining images to validation set as distractors. This operation turns the validation set into a large dataset of 92341 images, we denote this large dataset as MS-COCO2017, and the validation set as MS-val2017.

Visual Genome: This dataset consists of 108,077 images with detailed region-level text descriptions. Similarly, we select 15 groups images of complex scenes from this dataset, and choose 5 images from each group randomly and combine them together into a query set of 75 images.

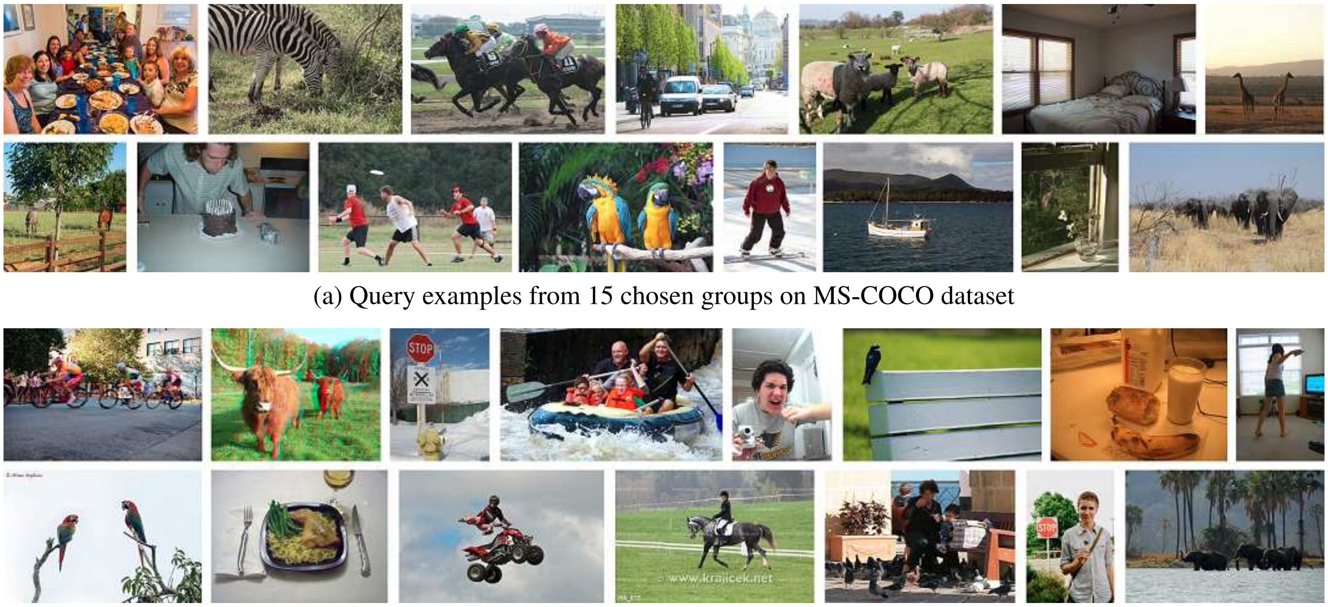
Fig.2 shows 15 query examples from each chosen group on MS-COCO dataset and Visual Genome dataset respectively.

3.3. Metrics

As in [16], we employ three standard metrics that are widely used in information retrieval tasks to evaluate the performance of all methods.

Normalized Discounted Cumulative Gain (NDCG): NDCG is a measure of ranking quality according to the accumulated relevance score of retrieved results. In our case, the relevance score of one image with respect to the query is the cosine similarity of their tf-idf representations. We compute the NDCG score for every image in query set and report their average value. Following [16] we compute the NDCG quality of top R results for different values of R to obtain the NDCG curves.

Spearman Rank Correlation: Spearman rank correlation assesses how well the relationship between two variables can be described using a monotonic function. Giving the ranking results returned by the proposed method and the ranking results returned by proxy tf-idf method for each query, it measures the correlation between these two rank-



(a) Query examples from 15 chosen groups on MS-COCO dataset

(b) Query examples from 15 chosen groups on Visual Genome dataset

Figure 2. Query examples on MS-COCO and Visual Genome.

ing results. Similarly, we report the average value of Spearman rank correlation for all queries.

Mean Average Precision(mAP): mAP is widely used in content-based image search to evaluate ranking results. When calculating the average precision in this experiment, either one image being relevant to the query or not is defined by either it belongs to the group of that query or not. The final mAP score is the mean value over all queries' average precision scores. We report the top R results for different values of R as in NDCG metric.

In the experiments, the NDCG score and Spearman rank correlation evaluate how well the proposed method correlates with the proxy tf-idf measure, while mAP evaluates how well it correlates with subjective cognition since the relevance is annotated by humans in this metric.

3.4. Results and Discussion

In this part, we show the image search results of the proposed method on three datasets: MS-val2017, MS-COCO2017 and Visual Genome.

Parameters. First of all, in order to determine the optimal values of two parameters in the proposed spatial-content image search method: α in Eq.(4) and β in Eq.(6), we test the retrieval performance on MS-val2017 when assigning different values to these parameters. Specifically, we study the impact of α by giving a fixed value to β, and after finding a promising value of α. We in turn study the impact of β with the chosen α value. We repeat this process until we find the best values for both α and β.

According to the experiment results, we choose $\alpha = 0.2$ and $\beta = 1$ for our method in this paper. Fig.3 shows the impacts of different parameter values on performance on MS-val2017 dataset. For the showed results, we set $R = 200$ in the NDCG and mAP measurements, and set the whole ranking list in the Spearman rank correlation. One can observe that when α changes from 0 to 1, the image search performance increases slowly at the beginning and then drops rapidly after $\alpha = 0.2$. Thus, the performance of $\alpha = 1$ is much worse than $\alpha = 0$, which indicates that searching by spatial-categories performs worse than searching by category labels and visual representations. This observation confirms the necessity for adding visual content information to spatial-semantic search.

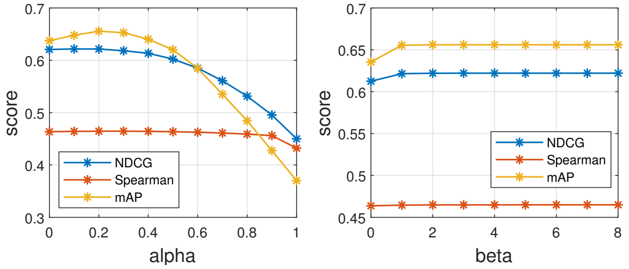


Figure 3. The impacts of the different parameters on image search performance on MS-val2017 dataset($R = 200$ for NDCG and mAP). Left: the impact of α on performance when $\beta = 1$. Right: the impact of β on performance when $\alpha = 0.2$.

As for β, there is an obvious improvement from $\beta = 0$ to $\beta = 1$, but the performance remains stable when β becomes larger. This means the penalization is beneficial to our method, however, a large β is not necessary as large β values leads to the value of $\widehat{S}$ in Eq.(6) approximating

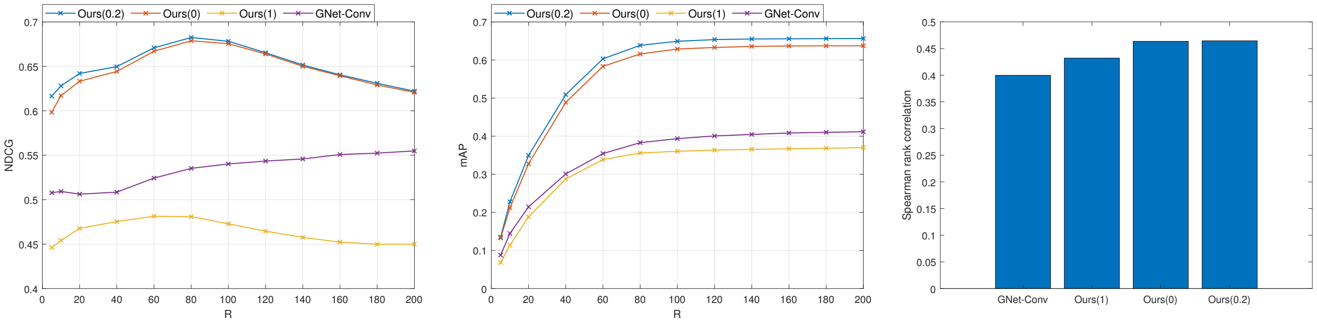


Figure 4. Performance of the proposed method and the baseline method on MS-val2017 dataset. From left to right: NDCG for different values of R, mean Average Precision for different values of R, Spearman Rank Correlation for whole ranking list.

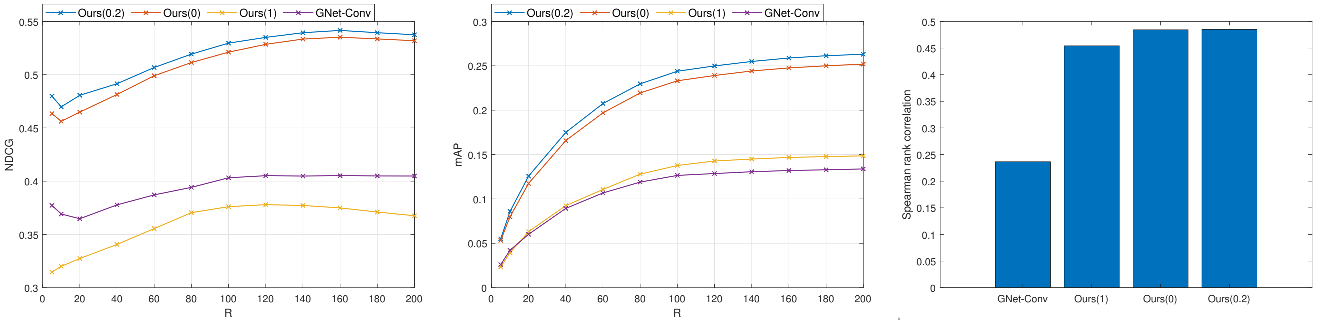


Figure 5. Performance of the proposed method and the baseline method on MS-COCO2017 dataset. From left to right: NDCG for different values of R, mean Average Precision for different values of R, Spearman Rank Correlation for whole ranking list.

to zero, and therefore make the final performance stable. Without losing the meaning of Eq.(6), we set $\beta \equiv 1$ in the following experiments.

Results. We present the performance of spatial-content image search method under three different values of parameter α: Ours(1) stands for the proposed method when $\alpha = 1$, it is an approximate spatial-semantic baseline method which uses spatial information and category labels for retrieval as in [16]; Ours(0) stands for the proposed method when $\alpha = 0$, it uses visual content information and category labels for retrieval; Ours(0.2) stands for the best condition of spatial-content search, which combines all information together and produces the most gratifying results.

Figs. 4, 5 and 6 compare the image search performance of the proposed method and the baseline method on MS-val2017, MS-COCO2017, Visual Genome datasets respectively. From these three figures, we have the following observations. On one hand, when $\alpha = 0$, the proposed method performs better than baseline GNet-Conv method with a significant gap on all three metrics. On the other hand, when $\alpha = 1$, the proposed method performs worse than GNet-Conv in terms of NDCG and mAP metrics in most cases, but with a slight gain over it using Spearman rank correlation. This again confirms the above discussion on parameter α. Meanwhile, the change from $\alpha = 0$ to $\alpha = 0.2$ further improves the search quality on three metrics, which proves the combination of spatial-semantic and visual content information is beneficial to image search in complex scenes.

Figs. 7 and 8 show the top retrieved images for a specific query on MS-COCO2017 and Visual Genome datasets respectively. Compared with the baseline method, the proposed method can capture the high-level semantic information within one image, such as vase and flower in Fig.7, and person, bench as well as birds in Fig.8, thus return more semantically similar results. Besides, when $\alpha = 0.2$, our method search images that are relevant to the query both spatial-semantically and visually.

Finally, we show quantitative comparison results of the evaluated methods in Table 1 with $R = 200$ for NDCG and mAP, and the whole ranking list for Spearman rank correlation. On the MS-val2017 dataset, the improvement of Ours($\alpha = 0.2$) over the baseline is 7% in terms of NDCG, 24% in mAP and 6% in Spearman rank correlation. On the MS-coco2017 dataset, the improvement of Ours($\alpha = 0.2$) over the baseline is more than 13% in NDCG as well as mAP, and 25% in Spearman rank correlation. On the Visual Genome dataset, the improvement gap is still significant, and is over 11% in terms of any metric. In a word, qualitative and quantitative results indicate that the proposed method leads to remarkable improvements in searching both visually and semantically relevant images.

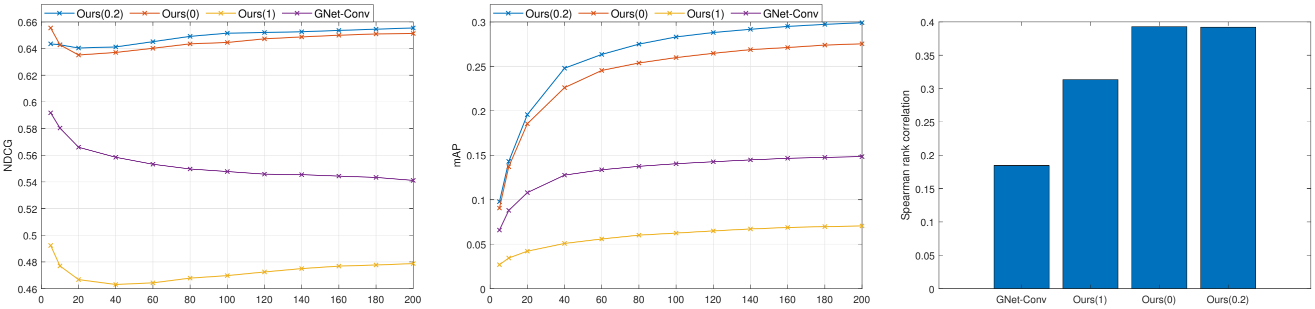


Figure 6. Performance of the proposed method and the baseline method on Visual Genome dataset. From left to right: NDCG for different values of R, mean Average Precision for different values of R, Spearman Rank Correlation for whole ranking list.

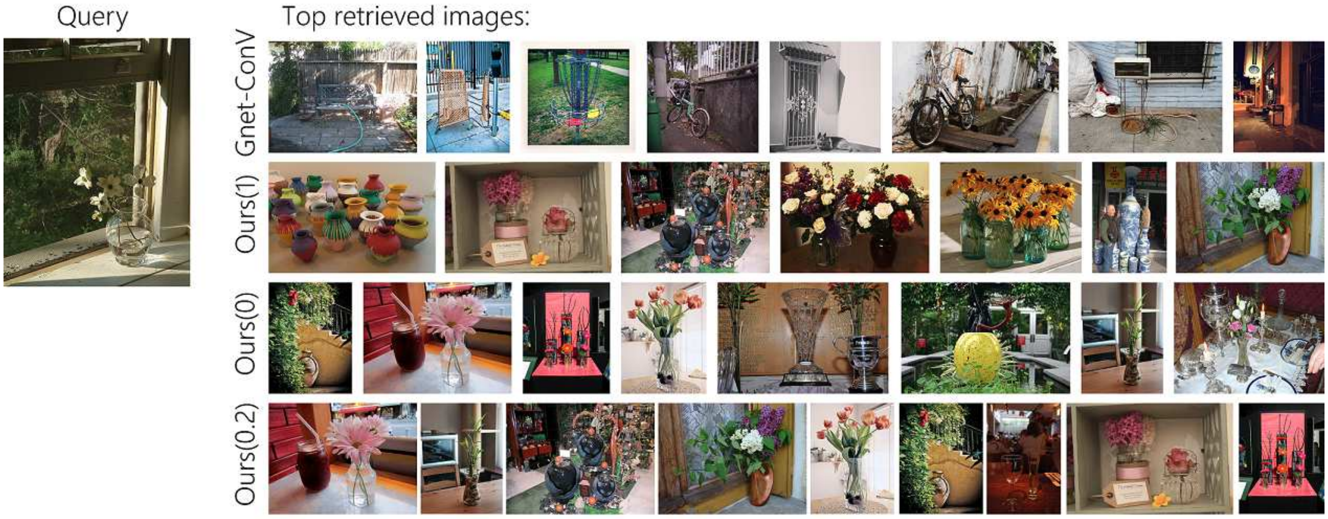


Figure 7. Qualitative results on MS-COCO dataset. Left column: query image. Right: from top row to bottom row: the top retrieved images of the methods GNet-Conv, Ours(1), Ours(0) and Ours(0.2).

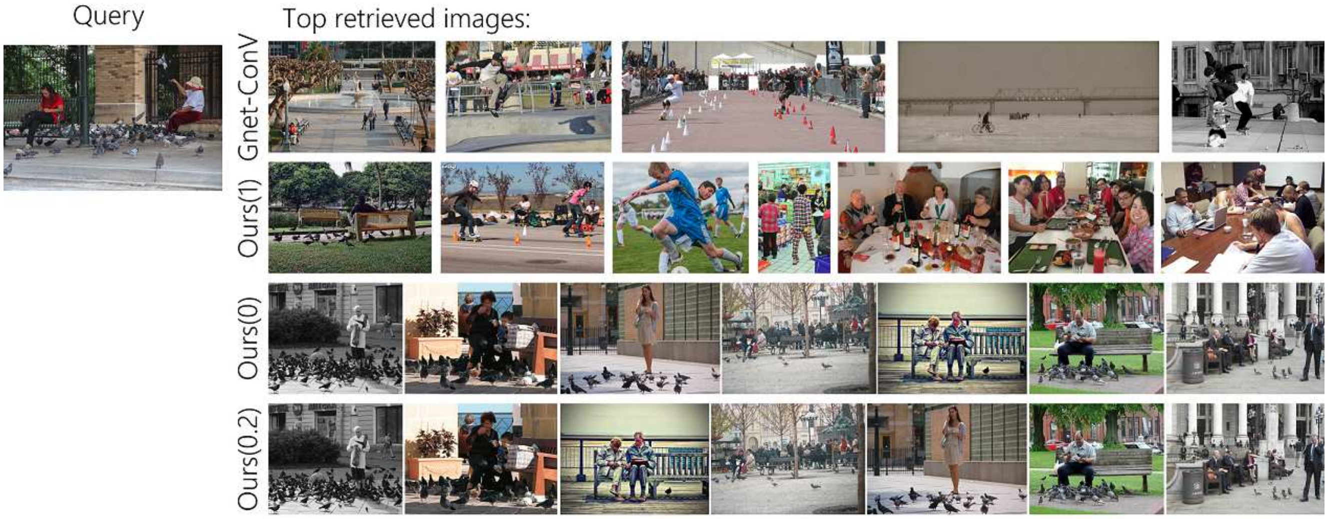


Figure 8. Qualitative results on Visual Genome dataset. Left column: query image. Right: from top row to bottom row: the top retrieved images of the methods GNet-Conv, Ours(1), Ours(0) and Ours(0.2).

Table 1. Comparison of different methods on three datasets ($R = 200$ for NDCG and mAP).

Method	NDCG	mAP	Spearman
MS-val2017			
GNet-Conv	0.5549	0.4116	0.3998
Ours($\alpha = 1$)	0.4500	0.3701	0.4321
Ours($\alpha = 0$)	0.6207	0.6374	0.4635
Ours($\alpha = 0.2$)	**0.6215**	**0.6555**	**0.4645**
MS-COCO2017			
GNet-Conv	0.4049	0.1338	0.2365
Ours($\alpha = 1$)	0.3676	0.1486	0.4542
Ours($\alpha = 0$)	0.5319	0.2519	0.4843
Ours($\alpha = 0.2$)	**0.5375**	**0.2630**	**0.4851**
Visual Genome			
GNet-Conv	0.5411	0.1485	0.1845
Ours($\alpha = 1$)	0.4787	0.0705	0.3134
Ours($\alpha = 0$)	0.6513	0.2754	0.**3929**
Ours($\alpha = 0.2$)	**0.6555**	**0.2991**	0.3920

3.5. Time complexity

Giving one query image, we first pass it to YOLOv3 and GoogleNet for extracting information that we need, and then construct the image representation with Eq.(2). We notice that these steps take about 80ms on a Tesla P100 GPU. Subsequently, we compare the query image with all database images either with Eq.(4) or with Eq.(6). In practice, we employ sequence comparison with 4 parallel threads, and the search time for each query is about 0.2s on MS-val2017, and about 5s on the two large-scale datasets MS-COCO2017 and Visual Genome. The time cost of our implementing can be further reduced by building inverted files with object category labels, however, this is beyond the scope of this paper and therefore we will not discuss it here.

4. Conclusion

In this paper, we present a solution for searching visually and semantically relevant images in complex scenes. Specifically, by leveraging recent advances on object detection, we use triplets of category labels, bounding boxes and deep visual features to jointly represent object instances contained in one image. With designed image representations, we then develop a novel similarity computation method that takes into consideration of both semantic and visual similarities to rank database images. Experiment results on MS-COCO and Visual Genome datasets show that the proposed spatial-content method improves the retrieval performance with a significant gap compared to the baseline method.

Acknowledgements

This paper was supported in part by the National Natural Science Foundation of China under Grant No. 61972312, and in part by the China Postdoctoral Science Foundation under Grant No. 2019M653335.

References

[1] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. Neural codes for image retrieval. In *ECCV*, pages 584–599, 2014. 1

[2] A. Bergamo, L. Torresani, and A. Fitzgibbon. Picodes: Learning a compact code for novel-category recognition. In *NIPS*, pages 2088–2096, 2011. 1

[3] I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115–147, 1987. 1

[4] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013. 1

[5] R. B. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 2

[6] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 2

[7] L. Gomez, Y. Patel, M. Rusinol, D. Karatzas, and C. V. Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *CVPR*, pages 2017–2026, 2017. 1

[8] A. Gordo and D. Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *CVPR*, pages 5272–5281, 2017. 1, 4

[9] A. Haque, A. Alahi, and L. Fei-Fei. Recurrent attention models for depth-based person identification. In *CVPR*, pages 1229–1238, 2016. 1

[10] K. He, G. Gkioxari, P. Dollar, and R. B. Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988, 2017. 2

[11] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, 2012. 1

[12] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 4

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014. 4

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*, pages 21–37, 2016. 2

[15] C. Lynch, K. Aryafar, and J. Attenberg. Images don't lie: Transferring deep visual semantic features to large-scale multimodal learning to rank. In *Proceedings of the 22nd*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 541–548, 2016. 1

[16] L. Mai, H. Jin, Z. Lin, C. Fang, J. Brandt, and F. Liu. Spatial-semantic image search by visual feature synthesis. In *CVPR*, pages 1121–1130, 2017. 1, 3, 4, 6

[17] V. Ordonez, X. Han, P. Kuznetsova, G. Kulkarni, M. Mitchell, K. Yamaguchi, K. Stratos, A. Goyal, J. Dodge, A. Mensch, et al. Large scale retrieval and generation of image descriptions. *International Journal of Computer Vision*, 119(1):46–59, 2016. 1

[18] S. Pang, J. Ma, J. Xue, J. Zhu, and V. Ordonez. Deep feature aggregation and image re-ranking with heat diffusion for image retrieval. *IEEE Transactions on Multimedia*, pages 1–11, 2018. 1

[19] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, pages 3384–3391, 2010. 1

[20] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, pages 1–8, 2007. 1

[21] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 2

[22] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. 2

[23] K. Rodden and K. R. Wood. How do people manage their digital photographs? In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 409–416, 2003. 1

[24] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun. OverFeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, pages 1–15, 2014. 2

[25] Sivic and Zisserman. Video Google: a text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003. 1

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 4

[27] R. Tao, A. W. Smeulders, and S.-F. Chang. Attributes and categories for generic instance search from one example. In *CVPR*, pages 177–186, 2015. 1

[28] G. Tolias, R. Sicre, and H. Jegou. Particular object retrieval with integral max-pooling of cnn activations. In *ICLR*, pages 1–11, 2016. 1

[29] L. Torresani, M. Szummer, and A. Fitzgibbon. Learning query-dependent prefilters for scalable image retrieval. In *CVPR*, pages 2615–2622, 2009. 1

[30] H. Xu, J. Wang, X. Hua, and S. Li. Image search by concept maps. In *SIGIR*, pages 275–282, 2010. 1

[31] A. B. Yandex and V. Lempitsky. Aggregating local deep features for image retrieval. In *ICCV*, pages 1269–1277, 2015. 1