Spatial-Content Image Search in Complex Scenes

Jin Ma[1], Shanmin Pang[1], Bo Yang[2], Jihua Zhu[1], Yaochen Li[1].

[1]Xi'an Jiaotong University, [2]Xi'an Polytechnic University.

ABSTRACT

The topic of image search has been heavily studied in the last two decades, but many works only focused on single-object search. In this work, we consider how to solve the problem of multi-objects search. Here we develop a novel 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.

Introduction to Image Search

The objective of image search is to return a ranked list of images that are relevant to a query within a very large database. A general image search framework is shown below in Fig.1.

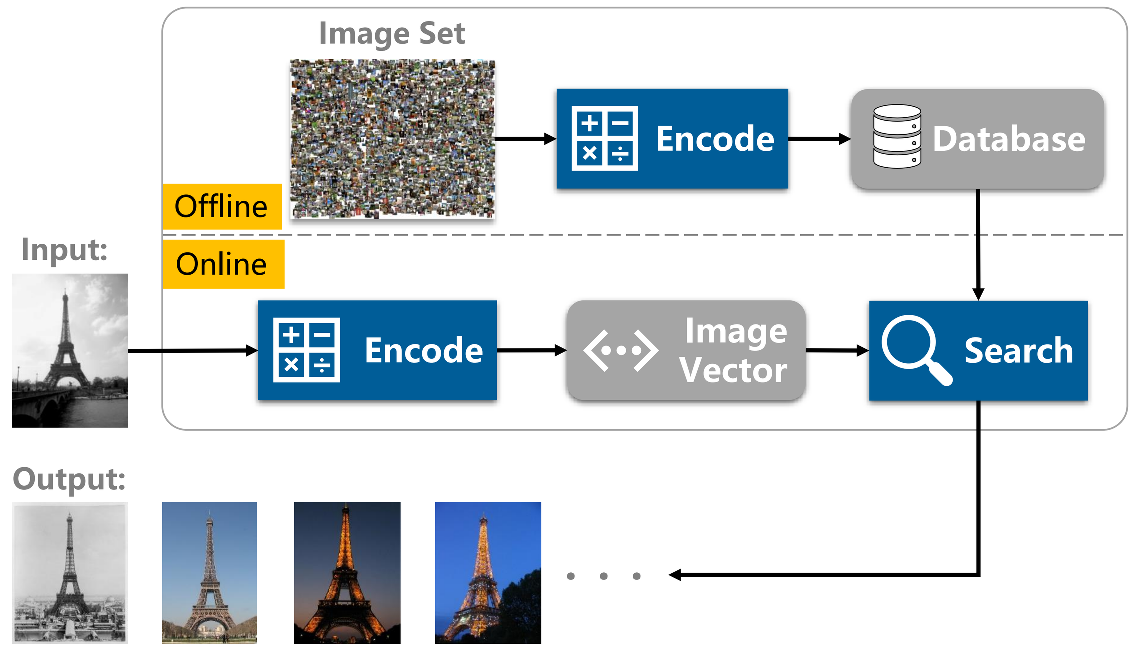

Fig.1 A general image search framework

Two key steps:

- Encode: construction of image representation.
- Search: computation of similarity score.

OBJECTIVES

- **Former works: single-object search.**

Typically, these methods use a vector in Euclidean space to represent an image. The similarity score between two images is defined by their vectors' L2-distance.

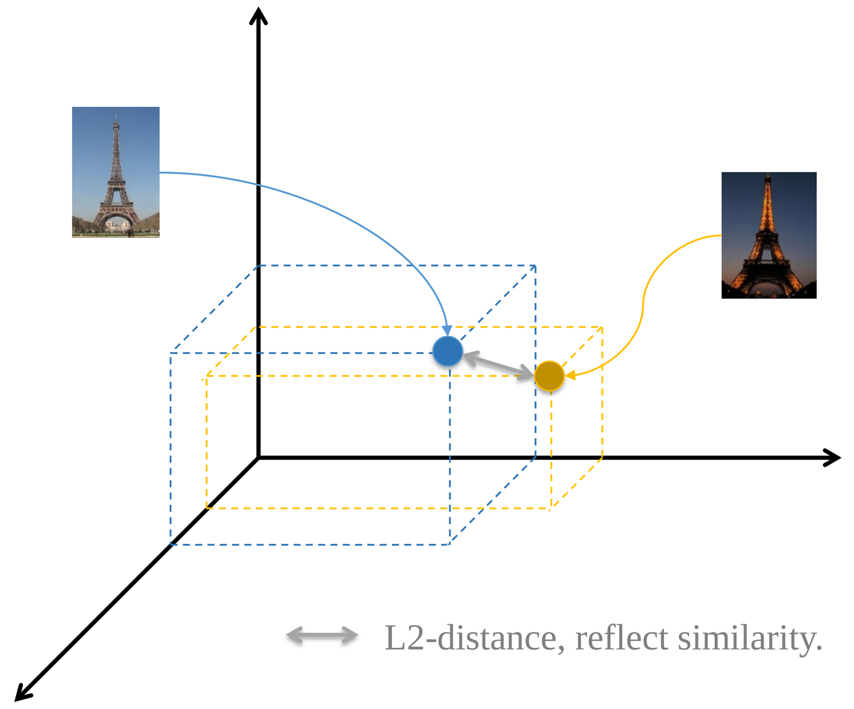

Fig.2 An illustration of the image representation and similarity computation in typical single-object search methods.

- **This work: multi-objects search.**

A single vector is insufficient for representing an image in complex scenes, since this kind of image usually contains multi-objects.

a. single object image b. multi-objects image

Fig.3 The difference between single-object and multi-objects image.

Two contributions:

- Design a new type of image representation.
- Customize corresponding similarity score.

METHOD

- **Image representation.**

We consider the following information of different objects when designing image representation:

- Visual content information.
- Spatial information.
- Semantic information.

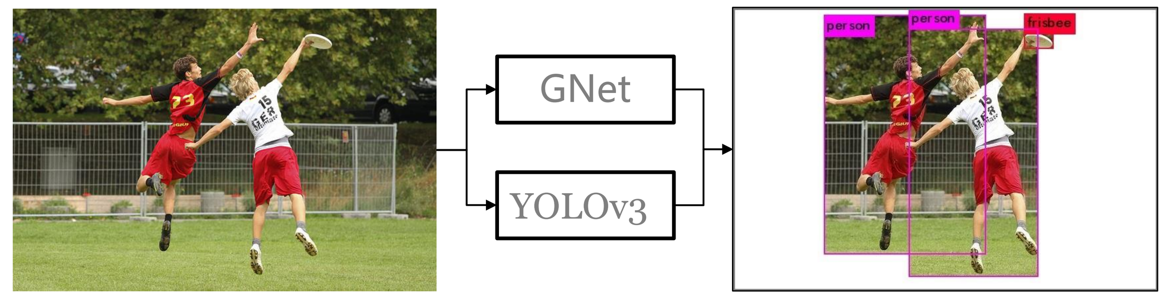

Fig.4 Information extracted from an image.

Then we represent an image $\boldsymbol{I}$ as:

$$\boldsymbol{I} = \{O_1, \dots, O_i, \dots, O_n\}, \text{ where } O_i = \{f_i,\ l_i,\ b_i\}.$$

- **Similarity Computation.**

The similarity between query $\boldsymbol{I^Q}$ and database image $\boldsymbol{I^D}$ is defined as:

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

The idea is: for every object O_i in query $\boldsymbol{I^Q}$, find its best match O_j in $\boldsymbol{I^D}$ and compute match score, Then S is the average of all match scores.

Fig.5 Matched objects (same color) between two different images.

EXPERIMENT

- Datasets: MS-COCO, Visual Genome.
- Standard relevance score:

  tf-idf BoW representation of captioned texts.
- Metrics: NDCG, Spearman, and mAP.
- Baseline: GoogLeNet convolutional features.

Table.1 Comparison to baseline method.

Method	NDCG	mAP	Spearman
	MS-COCO		
GNet-Conv	0.4049	0.1338	0.2365
Ours(best)	0.5375	0.2630	0.4851
	Visual Genome		
GNet-Conv	0.5411	0.1485	0.1845
Ours(best)	0.6555	0.2991	0.3920

Both quantitative and qualitative results indicate that the proposed method leads to remarkable improvements in searching both visually and semantically relevant images.

Query image:
(A bottle of flowers in front of the window):

Search with GNet-Conv features:

Search with proposed method:

Fig.6 An example of qualitative results on MS-COCO dataset.

CONCLUSION

When search images in complex scenes:

- Visual content information is insufficient, and some other information could be useful. Such as spatial, semantic, *et al*.
- Constructing appropriate image representation and designing corresponding similarity score are two key points.

REFERENCES

1. 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.
2. 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.
3. J. Redmon and A. Farhadi. Yolov3: An incremental improvement. CoRR, abs/1804.02767, 2018.
4. 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.

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.

FOR FURTHER INFORMATION

My email: majinwakeup@gmail.com

Video: https://www.youtube.com/watch?v=v9cO2KV6UmM

Poster: https://github.com/MaJinWakeUp/spatial-content/blob/master/Poster-final.pdf