

# Visual Knowledge Representation Learning

## A Survey Towards Briding the Semantic Gap by Connecting Language and Vision

Hanwang Zhang · Zhiyuan Liu · Tat-Seng Chua

Received: date / Accepted: date

**Abstract** Here, I simply describe some motivations of the title and the paper organization. You are more than welcomed to give comments.

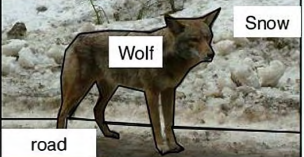


First, I think all of us might agree on the main title; for the subtitle, I use the most classic term “semantic gap” in multimedia to highlight the fundamental impact of this survey. By using “connecting language and vision”, which is the trending topic in both multimedia and computer vision, I would like to highlight that our survey is more focused on the recent progress on the marriage of NLP and CV.

Second, organizations are listed as section titles, followed by some informal descriptions of a section. Some seed references are also cited.

**Keywords** Semantic Gap · Knowledge Base · Representation Learning · Natural Language Processing · Computer Vision

### 1 Introduction

1. We recall the ever-lasting *Semantic Gap* in multimedia community, since the very well-known Smeulders’ survey paper [10]. Semantic gap states the challenge that the low-level visual features are usually unable to encode all the semantics needed for end tasks such as image retrieval and classification. 2. Our community,

<b>Semantics</b> <i>object relationships and more</i>	Wolf on Road with Snow on Roadside in Yosemite National Park, California on 24/1/2004 at 23:19:11GMT
<b>Object Labels</b> <i>symbolic names of objects</i>	
<b>Objects</b> <i>prototypical combinations of descriptors</i>	
<b>Descriptors</b> <i>feature-vectors</i>	Segmented blobs, Salient regions, Pixel-level histograms, Fourier descriptors, etc...
<b>Raw Media</b> <i>images</i>	

**Fig. 1** The Semantic Gap: Hierarchy of levels between the raw media and full semantics.

Hanwang Zhang  
School of Computing, National University of Singapore  
E-mail: hanwangzhangr@gmail.com

Zhiyuan Liu  
Department of Computer Science, Tsinghua University  
E-mail: liuzy@tsinghua.edu.cn

Tat-Seng Chua  
School of Computing, National University of Singapore  
E-mail: dcscts@nus.edu.sg

the main goal of multimedia analytics is to bridge this gap, benefiting any forms of multimedia information retrieval. Techniques that tackle this, can be categorized into the following three areas:

- **Visual Features.** Visual features represent an image as vectors, which are the most effective and efficient way for visual representations. However, it is

also vulnerable to variations in visual world, which result in large semantic gap. Fortunately, recent progress in deep learning has impressively robustify the features.

- **Semantic Concepts.** Visual features are not human interpretable and thus inevitably hinders the user-content interactions. To achieve this, building a set of semantic concepts is a promising way, which is widely used in TRECVID events. However, these semantics are unstructured and thus do not provide deep understanding of images.
- **Semantic Descriptions.** A full semantic description for an image such as objects, relations among objects and even the background knowledge of them, is our ultimate goal.

Figure 1 from [5] illustrates where the semantic resides in content understanding. With the development of CV, we can do well from the first level to the fourth level, however, the fifth level is still an open issue, which we claim that it should be resolved with a Visual Knowledge Base (VKB). And we believe the timing is ripe to do such research.

Talk about why we use a knowledge base for visual representation. Briefly introduce the recent progresses in NLP and CV that exploit knowledge base.

Paper organizations.

## 2 Visual Semantics

In this section, we briefly review our efforts made to bridge the semantic gaps in the past two decades. Generally, they are various of semantic classifiers trained for describing images.

The methods introduced in this section is generally based on classification paradigm. Compared to the rich reasoning that can go through a person’s mind upon seeing an object, a typical object classifier is doing a “shallow” reasoning and hence results in limited semantic understanding.

### 2.1 Concepts

High-level semantic concepts

### 2.2 Attributes

Intermediate-level concepts. These two sections can be found in my PhD thesis.

### 2.3 Relations

Such as actions, verbs, or other interactive semantics. These semantics are rarely investigated in CV and MM.

## 3 Knowledge Base

I think Zhiyuan can organize this.

## 4 Visual Knowledge Base

Inspired by the KB advances in NLP, the CV community has recently begun to explore Visual Knowledge Base (VKB).

Once again highlight the necessity of building VKB. For example, the incomplete VKB: ImageNet, has greatly pushed the development of CV in recent years.

### 4.1 Existing Organizations

Introduce some existing VKB organizations.

#### 4.1.1 Visual Data Population to Existing Knowledge Base

Based on existing KB, such as WordNet, we populate images according to the base. Such as the well-known ImageNet, and the two outdated projects: LSCOM and Vispedia. In fact, any visual datasets with labels can be considered as VKB of this kind.

#### 4.1.2 Image as Entities

The above KBs considered conventional concepts as entities. In [15], they consider each single image as an entity. But I think it is not a principled way.

#### 4.1.3 Others

### 4.2 Constructions

How to build a VKB. A good VKB should be **Large Scale** (this may require automatic techniques), **Well structured** and **Incremental** (easy for insertion and deletion). Zhiyuan may define in a more formal way of what a VKB should be like.

#### 4.2.1 Crowdsourcing

Invite human labelers to annotate, such as ImageNet, and the recent Visual Genome [7].

#### 4.2.2 Automatic Semantic Discovery

Discovering semantic concepts and hierarchies from Web data. This is the Webly-Discovered knowledge base. Most of them are merely noisy but automatically discovered datasets, such as NEIL [3,2], my MM14 paper [13], Microsoft ImageKB [12], LEVAN [4]. But few of them have rich relations among entities.

#### 4.2.3 Automatic Relation Verification

Discovering relations from web data. Such as mining object affordances [14,1] and relation verifications [11,8].

### 5 Visual Knowledge Representation Learning

Why learning a representation is necessary? In other words, we want to know how to use VKB in a computational way: how to ground images? how to predict semantic facts given an image using VKB? Yet, unfortunately, there is no principled way in VKB. This might be a chance for us to develop.

#### 5.1 Knowledge Grounding and Inference

Grounding entities in KB to visual regions (*e.g.*, objects) of images [6]. Almost all the above papers introduce somewhat ad-hoc methods. Formulations for inference. Given the visual information provided by the image (*i.e.*, after grounding), can we get the semantic description of the image as complete as possible?

#### 5.2 Learning Models

Formulations for training.

##### 5.2.1 Graphical Models

##### 5.2.2 Distributional Representations

### 6 Discussions and Future Directions

#### 6.1 Our proposals?

I feel that we should establish a principled framework for (a) VKB's construction: existing KB+automatic visual population, *e.g.*, mapping VisualGenome dataset [7] or ImageNet to Zhiyuan's KB; and (b) joint or partial learning the VKB, *e.g.*, although partial entities and

relations are annotated with images, how can we propagate the visual information to/ or exploit the rest KB to enhance the power of VKB?

In fact, in [14], it seems that it is a joint learning (every two entities have a weight to be learned) of visual evidence (*e.g.*, image visual features or its induced binary classifiers as weight and human pose estimations) and predefined rules in KB (*e.g.*, affordance edges, attribute assigning probabilities), using Markov Logic Network.

Interestingly, here is a recent scene graph parser that is compatible to the format of VisualGenome, called Stanford Scene Graph Parser [9]

#### 6.2 Visual Feature Enrichment

Use VKB to enrich existing visual features

#### 6.3 Visual QA

#### 6.4 Retrieval with Complex Multimedia Query

A query with text and image.

#### 6.5 Semantic Video Description

Video is a higher-order multimedia content, how to represent it in a KB.

#### 6.6 Incorporating Social Curation

Social intelligence can curate noisy but abundant visual informations for KB. How can we take advantage of it?

### References

1. Y.-W. Chao, Z. Wang, R. Mihalcea, and J. Deng. Mining semantic affordances of visual object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4259–4267, 2015.
2. X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1431–1439, 2015.
3. X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1409–1416, 2013.
4. S. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3277, 2014.

5. J. S. Hare, P. H. Lewis, P. G. Enser, and C. J. Sandom. Mind the gap: another look at the problem of the semantic gap in image retrieval. In *Electronic Imaging 2006*, pages 607309–607309. International Society for Optics and Photonics, 2006.
6. J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3668–3678. IEEE, 2015.
7. 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. *arXiv preprint arXiv:1602.07332*, 2016.
8. F. Sadeghi, S. K. Divvala, and A. Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1456–1464. IEEE, 2015.
9. S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 70–80, 2015.
10. A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *TPAMI*, 2000.
11. R. Vedantam, X. Lin, T. Batra, C. Lawrence Zitnick, and D. Parikh. Learning common sense through visual abstraction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2542–2550, 2015.
12. X.-J. Wang, Z. Xu, L. Zhang, C. Liu, and Y. Rui. Towards indexing representative images on the web. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1229–1238. ACM, 2012.
13. H. Zhang, Y. Yang, H. Luan, S. Yang, and T.-S. Chua. Start from scratch: Towards automatically identifying, modeling, and naming visual attributes. In *Proceedings of the ACM International Conference on Multimedia*, pages 187–196. ACM, 2014.
14. Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *Computer Vision–ECCV 2014*, pages 408–424. Springer, 2014.
15. Y. Zhu, C. Zhang, C. Ré, and L. Fei-Fei. Building a large-scale multimodal knowledge base system for answering visual queries. *arXiv preprint arXiv:1507.05670*, 2015.