3rd MMGR

ACM multimedia

Dublin, Ireland 27-31.10.2025

Prospective Analysis of Semantic Image Retrieval: Comparing Scene Graph, Visual Features, and Captions

Takahiro Komamizu
Nagoya University

https://taka-coma.pro/

Photographs

Drawings

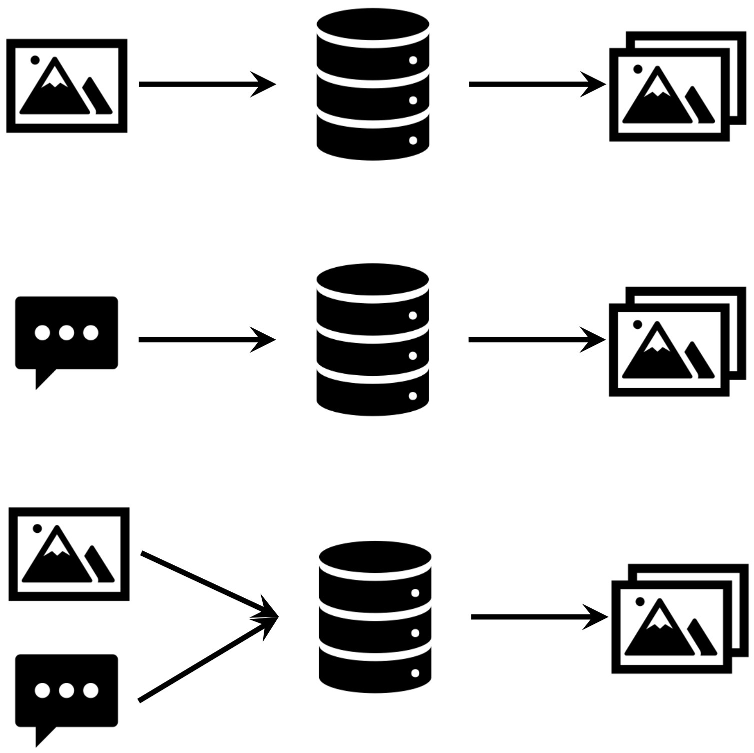
"A picture is worth a thousand words"

Diagrams

AI-generated

Image Retrieval: Search Images for Query Intent

- Content-based image retrieval (CBIR)
  - Query: an image containing target objects
- Text-based image retrieval
  - Query: a text describing what you want to find
- Compositional image retrieval
  - Query: a close image and
    a text describing desired modifications to it



The choice of retrieval models depends on
the expressiveness of query intent.

Possible Requirements on CBIR
➔ Various types of "relevance" ➔ **"Semantic" Image Retrieval**

Exact or Similar Image Search

- **Exact duplicate search:** Find identical copies (e.g., resized, cropped, or compressed versions).
- **Partial similarity search:** Find images containing overlapping or shared regions (e.g., same logo or object).
- **Geometrically invariant search:** Retrieve similar images regardless of rotation, scaling, or viewpoint changes.
- **Style similarity search:** Retrieve visually similar images in terms of color tone, composition, or texture.

Object- and Scene-Based Search

- **Object-based retrieval:** Find images that contain a specific object (e.g., "dog", "car").
- **Relational search:** Retrieve images with multiple objects in certain relationships (e.g., "a person walking a dog").
- **Scene-type search:** Retrieve scenes such as "beach", "classroom", or "city street".
- **Spatial relationship search:** Specify relations like "a cup on a table" or "a person standing beside a car".

Semantic or Concept-Level Search

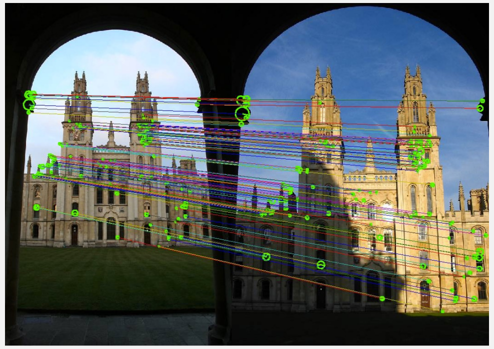
- **Semantic similarity search:** Find images conveying similar meanings (e.g., "wedding", "sports").
- **Emotion/mood-based search:** Retrieve images expressing certain moods (e.g., "bright atmosphere", "peaceful scenery").
- **Event or activity-based search:** Retrieve "people running", "people eating", etc.

Visual Feature / Style / Color-Based Search

- **Color-based retrieval:** Find images dominated by a certain color (e.g., red background).
- **Texture-based retrieval:** Search for similar texture patterns (e.g., fabrics, materials).
- **Shape-based retrieval:** Retrieve objects with similar outlines or silhouettes.
- **Art-style search:** Retrieve images with similar artistic styles (e.g., impressionist, sketch).

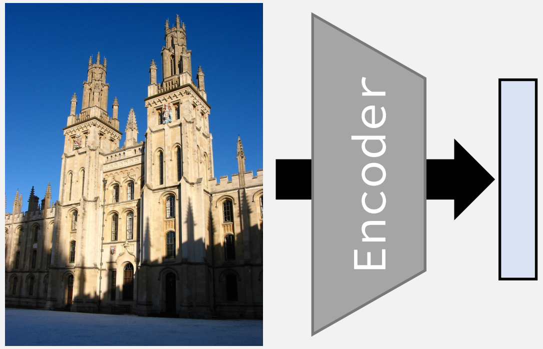
Representations in Semantic Image Retrieval

Low-level features:
SIFT, color histogram



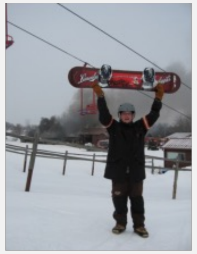
Appearances of **objects** matters.

Visual features:
DNN-based encoders



Appearance of **image** matters.

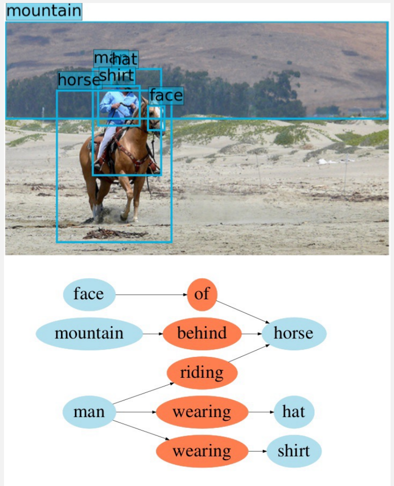
Caption:
textual description



The image depicts a person standing on a snowy terrain, likely a ski resort or a snowboarding area. The individual is wearing a black jacket, brown pants, and a helmet, which suggests they are engaged in winter sports, specifically snowboarding. The person is ...

Appearance **somewhat** matters.

Scene graph:
object-object relationships



Appearance **does not** matter.

The choice of representations depends on the requirements of query intent.

Scope of this research

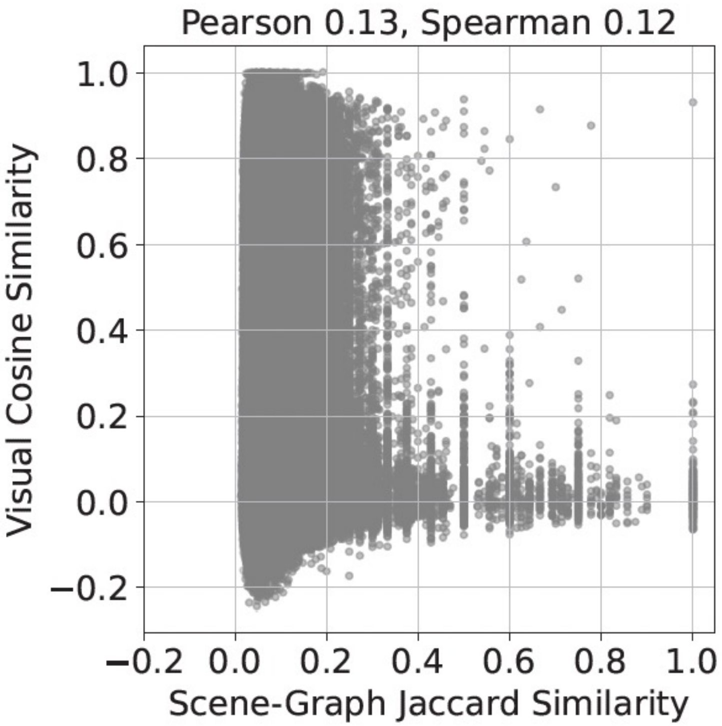
- Research objective: adaptive semantic image retrieval system
  - To deal with various requirements in CBIR.
  - No retrieval method may fit all requirements.
    - Because they are dependent on basic representations.
- This paper: analysis on relationships between representations
  - Dataset: Visual Genome[10]
  - Representations: Scene graph, Visual feature, Caption

Representations	Vectorization	Similarity Function
Scene graph	Annotation in the dataset	Jaccard similarity between sets of triplets
Visual feature	DINOv2[17]	Cosine similarity
Caption	Detailed caption generated by Qwen2-VL[25] and E5[24] to encode texts into vectors	Cosine similarity

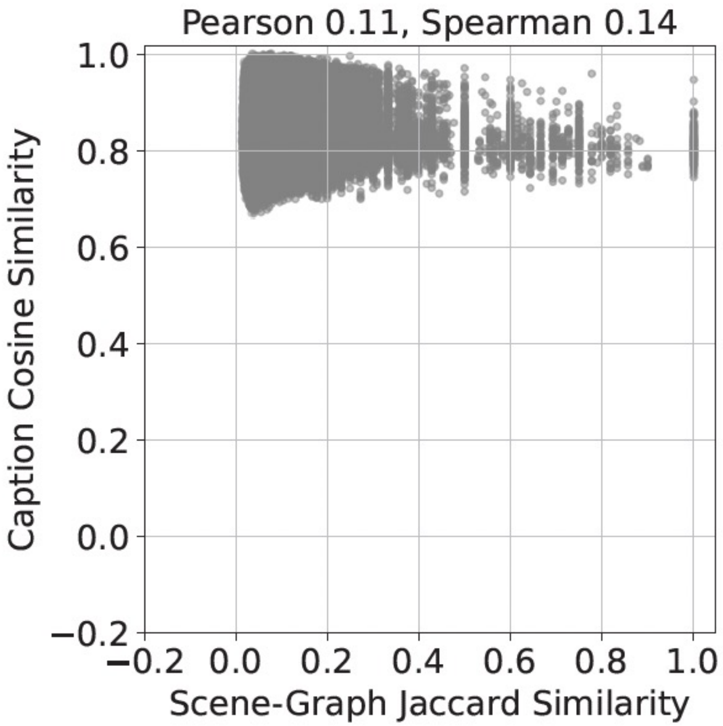
Correlation Analysis

Lower Correlation

Scene graph vs. visual and caption



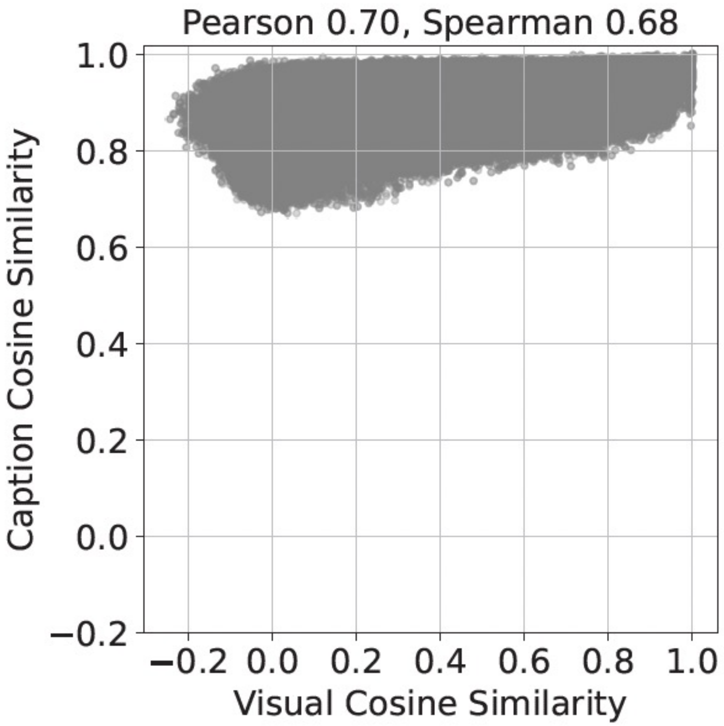
(a) Scene Graph vs. Visual Feature



(b) Scene Graph vs. Caption Embedding

Higher Correlation

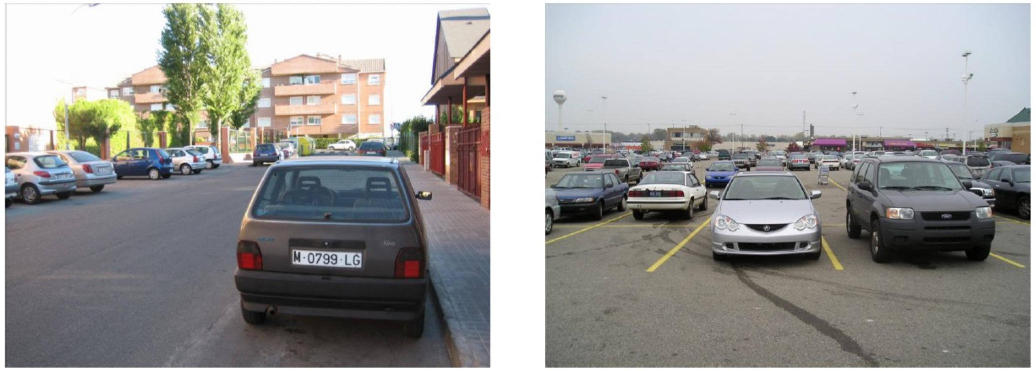
Visual vs. Caption



(c) Visual Feature vs. Caption Embedding

Examples of retrieval results

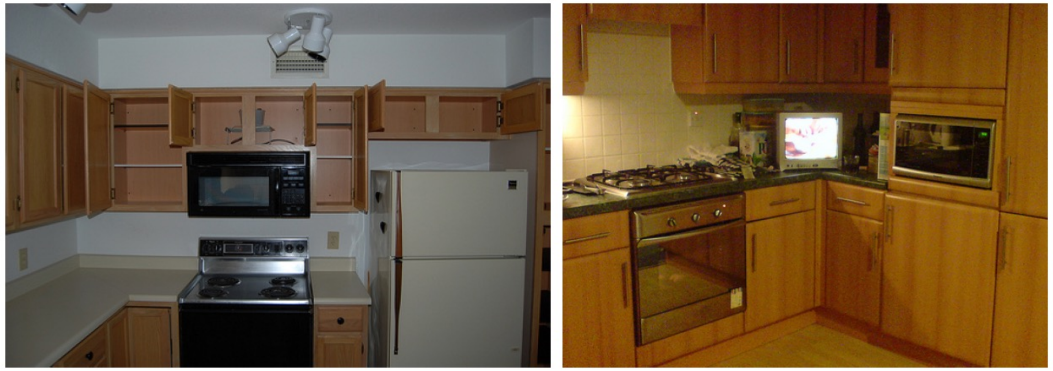
QE1 – SG: 0.75 (High), Vis: 0.13 (Low), Cap: 0.86 (Med.)



(a) Query

(b) Retrieved

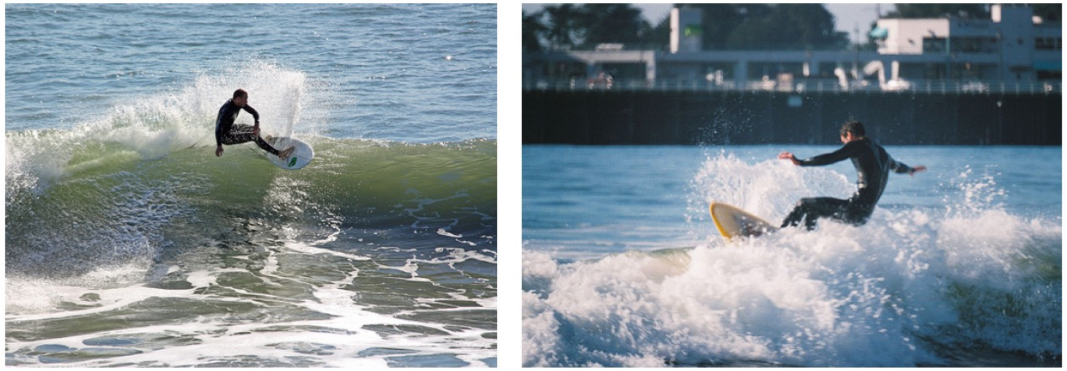
QE2 – SG: 0.03 (Low), Vis: 0.13 (Low), Cap: 0.98 (High)



(c) Query

(d) Retrieved

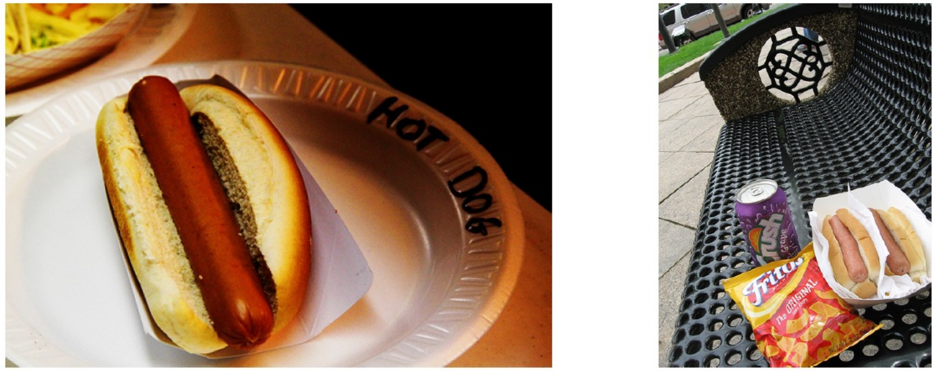
QE3 – SG: 0.67 (Med.), Vis: 0.91 (High), Cap: 0.92 (Med.)



(e) Query

(f) Retrieved

QE4 – SG: 0.04 (Low), Vis: 0.81 (High), Cap: 0.79 (Low)



(g) Query

(h) Retrieved

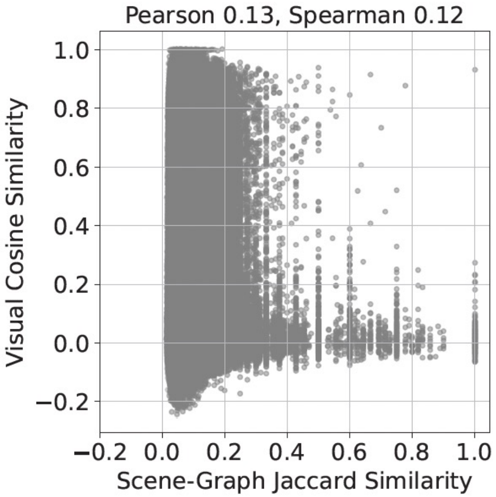
(SG: Scene Graph, Vis: Visual Feature, Cap: Caption)

Lessons Learned

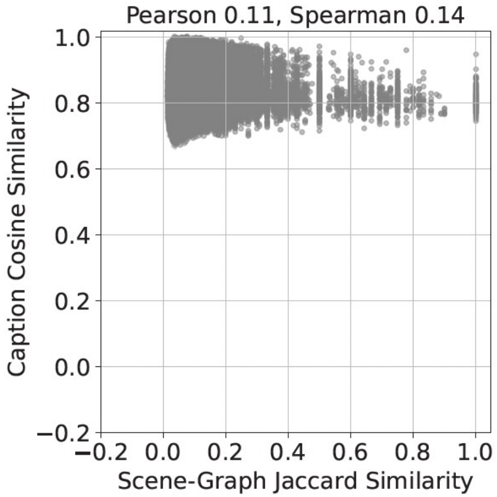
- Weak correlation b/w scene-graph and embedding while high correlation b/w embedding methods
  ➔ Scene-graphs capture distinct relational semantics.
  - Suggestions
    - Scene-graph: Applications (e.g., robotics or surveillance) object configurations matter may benefit from incorporating scene-graph information.
    - Visual and caption: They are better suited for aesthetic or thematic retrieval tasks where global appearance and scene atmosphere dominate.
- To realize adaptive semantic image retrieval system
  - Investigation of fusion strategies
    - weighted combination, learnable fusion, re-ranking with structural constraints

Conclusion

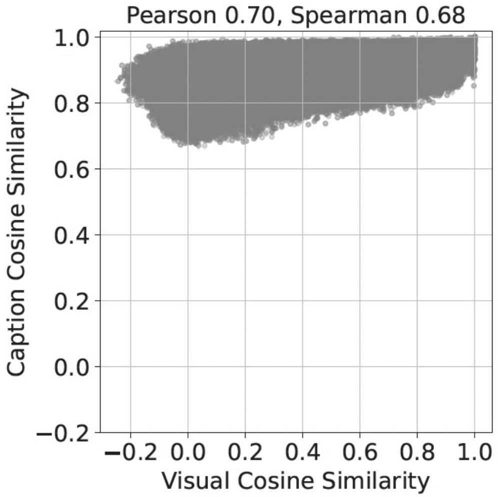
- Analysis on relationships between representations
  - Dataset: Visual Genome[10]
  - Representations: Scene graph, Visual feature, Caption
  - Results: Weak correlation b/w scene-graph and embedding while high correlation b/w embedding methods



(a) Scene Graph vs. Visual Feature



(b) Scene Graph vs. Caption Embedding



(c) Visual Feature vs. Caption Embedding

QE2 – SG: 0.03 (Low), Vis: 0.13 (Low), Cap: 0.98 (High)

(c) Query

(d) Retrieved

QE4 – SG: 0.04 (Low), Vis: 0.81 (High), Cap: 0.79 (Low)

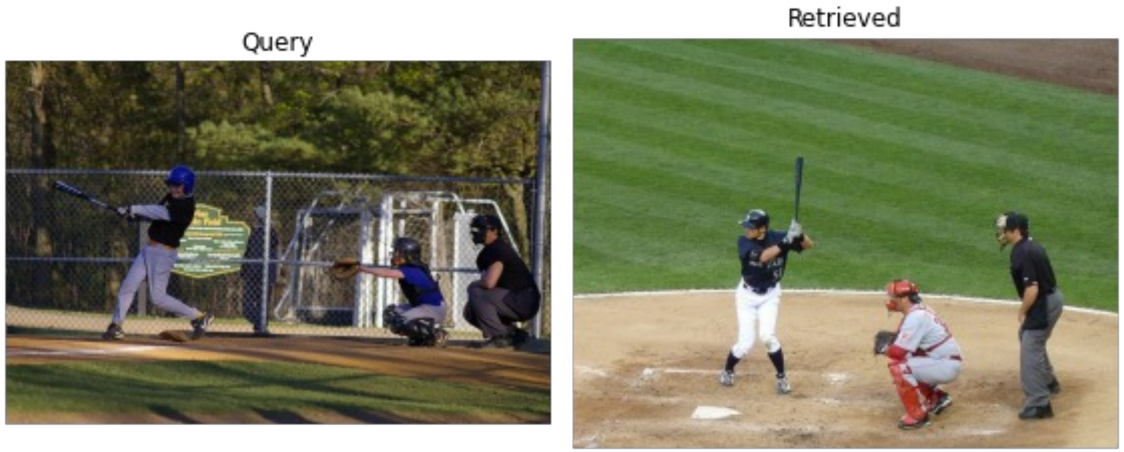
(g) Query

(h) Retrieved

Future directions

- Increase the number of datasets to analyze
- Improve scene-graph quality
  - Develop open-vocabulary and detailed scene-graph generation method
- Realize an adaptive semantic image retrieval mechanism
  - Estimation of query intent is a core issue.

SG: 0.07, Vis: 0.16, Cap: 0.98



? what users want with the input image?

Exact or Similar Image Search

Object- and Scene-Based Search

Semantic or Concept-Level Search

Visual Feature / Style / Color-Based Search