Sensory Gap: Why You Shouldn't Overlook It

Reza Bahmanyar[1,2], Ambar Murillo Montes de Oca[1]
[1]Remote Sensing Technology Institute, German Aerospace Center (DLR), Münchener Str. 20, D-82234 Weßling, Germany
[2] Munich Aerospace Faculty, Munich, Germany

Sensory & Semantic Gaps

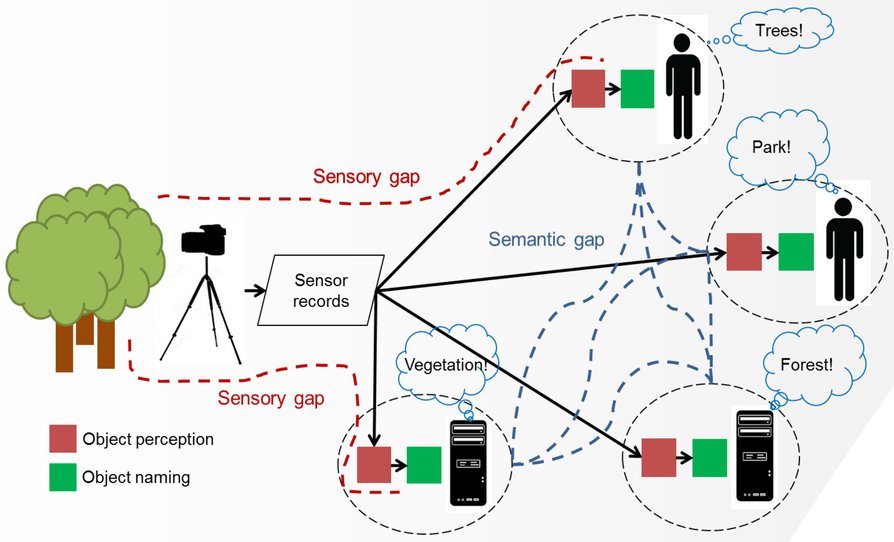


Fig. 1: Depiction of the sensory & semantic gaps, & their relationship

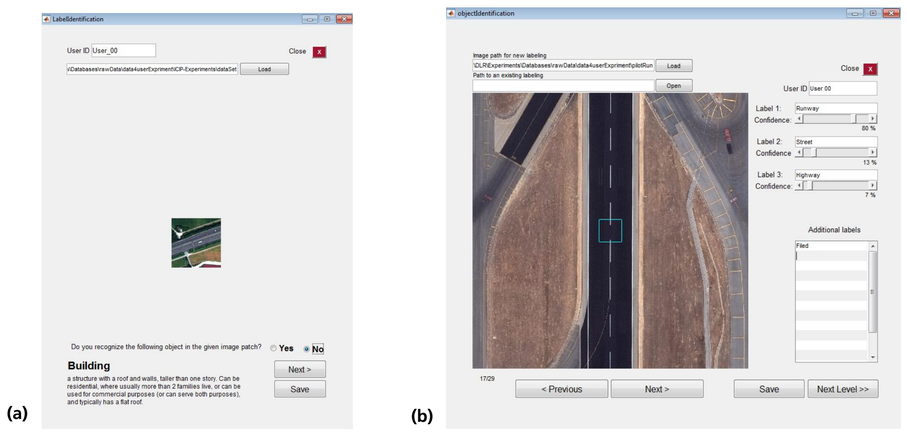
The **sensory gap** has been defined as the gap between a real life scene, and the information of this scene captured by sensors [1]. A specific object can either be perceived directly by the user, or detected after processing the information in a machine learning application, which is highly dependent on the feature descriptors being used (e.g., Scale Invariant Feature Transform). In either case, a sensory gap exists. Once the user perceives the object, a more conceptual task takes place: object naming. This is where the **semantic gap** plays a larger role. The semantic gap has typically been defined as the difference between the user's understanding of objects in an image and the computer's interpretation of those objects [1]. In addition to this gap between the user and computer, we must consider that each user will interpret images differently and use different terms to label the objects within them. Therefore, there will be a semantic gap between different users. This is what we call the "linguistic semantic gap" [2]. Additionally, due to the effects of the sensory gap (e.g., which feature descriptors are used), and the applied learning algorithm, each image mining system will identify different objects in an image. This results in a gap between different systems' object identification. Since discrimination, which is mostly affected by the sensory gap, is a basic step for object identification, the sensory gap has a clear effect on the semantic gap.

Why Do We Study the Sensory Gap?

- In Earth Observation (EO), image product properties are fixed. Different tasks require different image properties, and studying the sensory gap helps:
  - find appropriate image properties for annotation and learning tasks, such as optimal image patch size
  - find the best combination of data sources from different sensors (SAR, multi-spectral), with different properties (e.g., resolution)
  - train image analysts and annotators how to account for the sensory gap when identifying objects
- In addition to working with a specific patch size, analysts can be trained to understand how objects of interest look when the image is taken from a bird's eye view by different sensors, and how to identify hard-to-classify objects by using different contextual clues to gather missing information (such as height, depth, and material).
- The sensory gap from the computer side is affected by the feature descriptors being used. Object discrimination, which is mostly affected by the sensory gap, is a basic step for object identification; therefore the sensory gap has an effect on the semantic gap. The semantic gap is a measure of the relevance of the information provided by the computer for the user. Therefore, studying the sensory gap provides a way to identify feature descriptors which present the most relevant information for the specific task.

Current & Future Research

- Research on the interaction between the causes of the sensory gap should be extended
- The relationships between the sensory and semantic gaps should be further studied
  - Considering the sensory gap is necessary for semantic gap assessment
- We are currently studying the effects of patch sizes on the sensory gap. In EO, the Field of View (FoV) of image products is fixed, however recent research has focused on patch-wise image analysis, as opposed to analyzing the whole scene. Patch size affects the amount of content present, limiting the contextual clues available. The results of an analysis conducted solely on patches may differ from the results derived from a complete scene. This shows the effect of patch size on image interpretation.
- Fig. 2 shows two tools designed for our user experiments exploring the effects of contextual clues on object identification. The "Label Identification Tool" (Fig. 2.a) focuses the user on finding a specific object in a patch, whereas the "Object Identification Tool" (Fig. 2.b) asks the user to label multiple objects and state their confidence level in their labels.



(a) (b)

Fig. 2: (a) Screenshot from our Label Identification Tool **(b)** Screenshot of our Object Identification Tool

Causes Behind the Sensory Gap

Causes behind the sensory gap can lie at the scene (e.g., clutter, occlusion) or sensor levels (e.g., perspective, resolution). In Earth Observation (EO), the sensory gap is rather wide due to sensors which record visual information differently from the human visual system [3]. The sensory gap is affected by the complexities of the images used in EO, such as the resolution, perspective, or scale of the visual information [4].

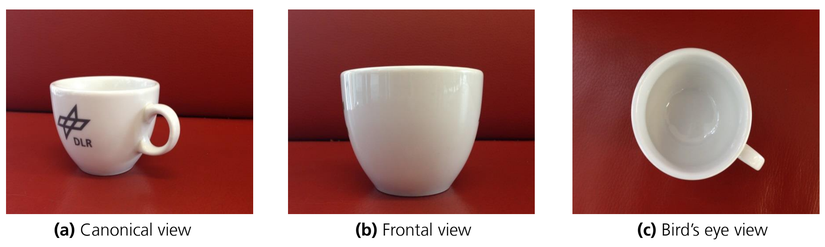
- **Sensor:** Different sensors produce various types of data with distinct properties.

(a) **(b)**

Fig. 3: (a) WorldView-2 satellite image of Munich, Germany **(b)** TerraSAR-X satellite image of the same area

For example, multi-spectral sensors capture images at various frequency bands across the electromagnetic spectrum, some of which are not visible to the human eye (Fig.3.a). As another example, SAR is a form of radar which transmits pulses of radio waves, and records the amplitude and phase information of the received echoes (Fig.3.b). This active form of signal recording stands in contrast to the passive form of the human visual system.

- **Perspective:** Humans are not accustomed to a bird's eye view.



(a) Canonical view **(b)** Frontal view **(c)** Bird's eye view

Fig. 4: A cup viewed from different perspectives

As described in the "Recognition by Components" theory [5], objects can be segmented into their geometric components ("geons"), and humans recognize them based on the identification of their geons and their structural relationships, which are then matched to mental representations. For humans, it is easiest to identify objects from a "canonical perspective", such as in Fig. 4.a. Here depth information is present, and the structural relationships between geons are clear. When objects are viewed from a completely frontal perspective (Fig. 4.b), depth information is occluded, along with certain structural relationships. A view from above (Fig. 4.c) loses depth information. Instead of a cup made up of a cylinder and a semi-circular shape, we now see a circle and a small rectangle. Object recognition should be perspective invariant, as long as the structural relationship between geons remains identifiable. This is not the case when objects are viewed from above, since major object components can be occluded, making it harder to match the object to the stored mental descriptions. Therefore, from this perspective, object identification is more difficult [5].

- **Resolution:** It affects visible details and the clarity of object contours.

In an image with many relatively small objects present at a relatively high density (such as in Fig. 5.a), it is clear how a higher resolution would aid in identifying object contours and attributes, and therefore helps in object identification.

- **Scale:** Determining object size can be problematic without a clear scale.

In Fig. 5.a it is hard to estimate the size of the depicted rectangular objects. Are they trucks on a road? Train wagons? Or buildings? No clear scale in Fig. 5.b can lead the user to wonder if they are viewing a pool at a sport club or perhaps a rooftop.

- **Image Patch Size:** Enlarging the image patch size increases the amount of contextual information available. Context is important because it provides information on spatial relations, semantic associations, and global scene properties.

Fig. 5.c includes the same objects as in Figs. 5.a and b, however the user has much more contextual information surrounding the objects, which aid in their identification. The "pool at a sport club" is actually a rooftop structure on a shopping center. What looked like trucks on a road can now be identified as buildings based on their numbers and positions relative to the shopping center.

(a)

(b)

(c)

Fig. 5: (a) Trucks or buildings? **(b)** Pool or rooftop? **(c)** A bigger patch size provides additional context information

References

[1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.
[2] R. Bahmanyar, A. Murillo Montes de Oca ,M. Datcu, "The Semantic Gap: An Exploration of User and Computer Perspectives in Earth Observation Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 10, pp. 2046-2050, 2015.
[3] R. Bahmanyar and M. Datcu, "Measuring the semantic gap based on a communication channel model," in *Proc. 20th IEEE International Conference on Image Processing (ICIP)*, Melbourne, Australia, 2013, pp. 4377–4381.
[4] A. Murillo Montes de Oca, N. Nistor, and M. Datcu, "Creating a Reference Data Set for Satellite Image Content Based Retrieval," in *Proc. Conference on Big Data from Space (BiDS)*, Frascati, Italy, 2014, pp. 71–75.
[5] I. Biederman, "Recognition-by-components: A theory of human image understanding," *Psychological Review*, vol. 94, no.2, pp. 115–147, 1987.

REMOTE SENSING TECHNOLOGY INSTITUTE