EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

A Hybrid Framework for Visual Positioning: Combining Convolutional Neural Networks with Ontologies

Abdolreza Mosaddegh [1,*], Sérgio Lopes[2], Habib Rostami[3], Ahmad Keshavarz[3] and Sara Paiva[2]

[1]King's College London, London, United Kingdom
[2]Instituto Politécnico de Viana do Castelo, Viana do Castelo, Portugal
[3]Persian Gulf University, Bushehr, Iran.

Abstract

Visual positioning is a new generation positioning technique which has been developed rapidly during recent years for many applications such as robotics, self-driving vehicles and positioning for visually impaired people due to advent of powerful image processing methods, especially Convolutional Neural Networks. Nowadays, deep Convolutional Neural Networks are capable of classifying images with high accuracy rates; however, comparing visual perception by a human being, pure Neural Networks lack background knowledge which is essential for estimating the position through a reasoning process. In this paper we present a hybrid framework for employing ontologies over Convolutional Neural Networks to integrate a knowledge-based reasoning with Neural Networks for taking advantages of capabilities similar to human brain's functions. The proposed framework is generic so it can be applied to a wide variety of scenarios in smart cities where visual positioning represents added value.

Keywords: Visual Positioning, Convolutional Neural Network, Ontology, Image Understanding

Received on 12 November 2022, accepted on 20 December 2022, published on 29 December 2022

doi: 10.4108/ew.v9i40.2959

1. Introduction

Positioning is a mechanism for determining the position of an object in space (Gu et al., 2009). Several methods have been used in positioning, that raise from signal analysis to image processing, presenting distinct coverages and accuracies. One of the most famous and successful positioning systems is GPS (Global Positioning Systems); however, it has its own limitations. The precision of the system may be affected by poor signal in some locations such as near tall buildings or during adverse weather conditions. On the other hand, some applications, especially in dense urban areas, require more precision. This need led to the introduction of new positioning methods such as Visual Positioning (VP) which is capable of providing increased accuracy, as an alternative or a complement to GPS.

VP is a sub-class of positioning focused on computer vision methods to estimate the user's location. The main applications of VP include positioning for visually impaired people, robotics and self-driving vehicles, among other applications and domains. VP methods usually employ a dataset of geo-tagged images, which contains images together with their location (Figure 1). The reference images are used to train a model that can be used to recognize locations in the taken photos. The target position can be estimated using location of geo-tagged images and by considering the direction and angle of the imaging device.

*Corresponding author. Abdolreza.Mosaddegh@kcl.ac.uk

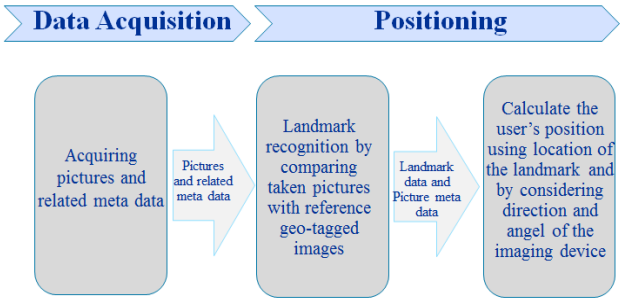

Figure 1. Traditional approach of visual positioning.

There are many studies on visual positioning in the literature focused on image comparison methods, especially for indoor positioning and navigation. Most of these studies use feature extraction techniques (e.g., SURF and SIFT) and Convolutional Neural Networks (CNN) to compare the photos with geo-tagged images. Rahman et al. (2019) developed a visual positioning and navigation system for smartphones. The images captured by smartphones are compared with pre-existing images in a dataset estimate the user position. The system can also identify obstacles in images that can be used for assisting visually impaired people.

In some studies, the positioning is a part of the Simultaneous Localization and Mapping (SLAM) technique. In SLAM, a map of the environment is produced and then the target position is derived (Aulinas et al., 2008). Deretey et al. (2015) proposed a method for indoor positioning using features of images to produce the structure of a scene. Features of reference images are stored in a dataset to compare with feature of taken photos for recognizing the location. The results have shown an accuracy of 10 mm for positioning in an indoor environment.

Li et al. (2019) also introduced an indoor positioning and navigation system for visually impaired people using the SLAM technique. When the user moves in the building, the system gradually produces an abstract map and computes the user position.

Visual positioning also has many applications in robotics and self-driving vehicles. Kostavelis and Gasteratos (2015) developed a semantic positioning and navigation system for robots. The robot classifies and labels all locations in an indoor environment and then estimates its position. In this approach, places are categorized into semantic locations such as bedroom, kitchen, etc. In a vision-based positioning for vehicles (Chen et al., 2016), a pool of models has been used for different scenarios, moments of day and weather conditions. The best model for an input image is selected and then geo-tagged images are used for calculating the position of the vehicle.

Since most studies in the existing literature are focused on improving visual aspects, interdisciplinary methods on using knowledge base integration with image processing techniques can provide new facilities for visual positioning to take advantage of semantic understanding besides deep learning capabilities. In this regard, and as the main contribution of this work, a novel framework is proposed to integrate the concept of ontology with image processing techniques towards a generic visual positioning system. This is a generic framework that can be applied to a wide range of applications in smart cities from robotics to positioning for visually impaired people.

In this paper, firstly we describe the concept of image understanding and then a conceptual framework to employ such capabilities in visual positioning is presented. Finally, we discuss the application of the framework in an urban environment.

2. From image classification to image understanding

Nowadays, many scholars believe that CNNs are the most efficient method for image classification (e.g. Liu et al., 2020) and during recent years many improvements have been made by Deep Neural Networks (DNN) which provide high accuracy to recognize visual objects (e.g. Li et al., 2017).

Neural Networks simulate brain functions for processing information (Fan et al., 2020) and in fact, the multilayer structure of the brain inspired the development of deep Neural Networks. However, the brain also has cognitive capabilities, which rely on the background knowledge and contains concepts and relations that can be used for reasoning. Although the visual signals are processed by neurons of the visual cortex of the brain, the brain also takes advantage of a working memory, which keeps the concepts required for image understanding. Pre-frontal cortex in the brain is a key part of this memory (Goldman-Rakic, 1990). Without such background knowledge, interpreting visual objects has many shortcomings, which cannot be addressed by pure neural networks. For example, it is completely a different understanding for a person who sees a distant object in the wildlife with features like a lion comparing the understanding by another person who sees an object with same features in a town square. Since there are associations between concepts with their locations in our brain, we assume that the first object is likely to be a real lion and the second one could be a sculpture of a lion though the visual features are the same. In contrast, a pure convolutional neural network approach classifies both objects as lion (Figure 2).

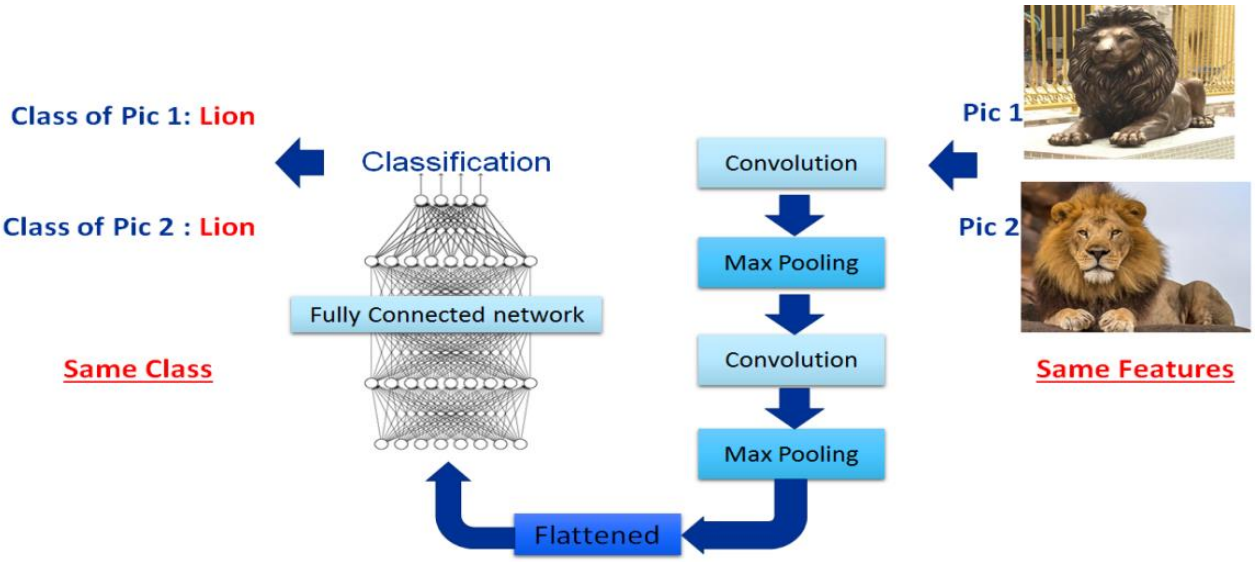

Figure 2. Same features lead to recognize a same class by a pure Neural Network approach.

Most studies on computer vision have focused on the improvement of quantitative techniques for image processing (e.g., Liu et al., 2020) and only a few of them (e.g. Gupta et al., 2015; Monroy et al., 2019) have employed knowledge-based approaches within artificial intelligence to integrate context knowledge with deep learning techniques.

There are different knowledge-based methods in the literature; however, one of the most effective tools to present formal descriptions of knowledge is ontology. An ontology describes a knowledge model in a specific domain with an automated processable language (Studer et al., 1998).

To make neural networks wiser and more similar to the human brain, the ontology provides various types of associations between concepts, which could be rules or possibilities that can be used for logical or probabilistic reasoning. A simple example of employing an ontology to simulate reasoning in the brain is shown in Figure 3.

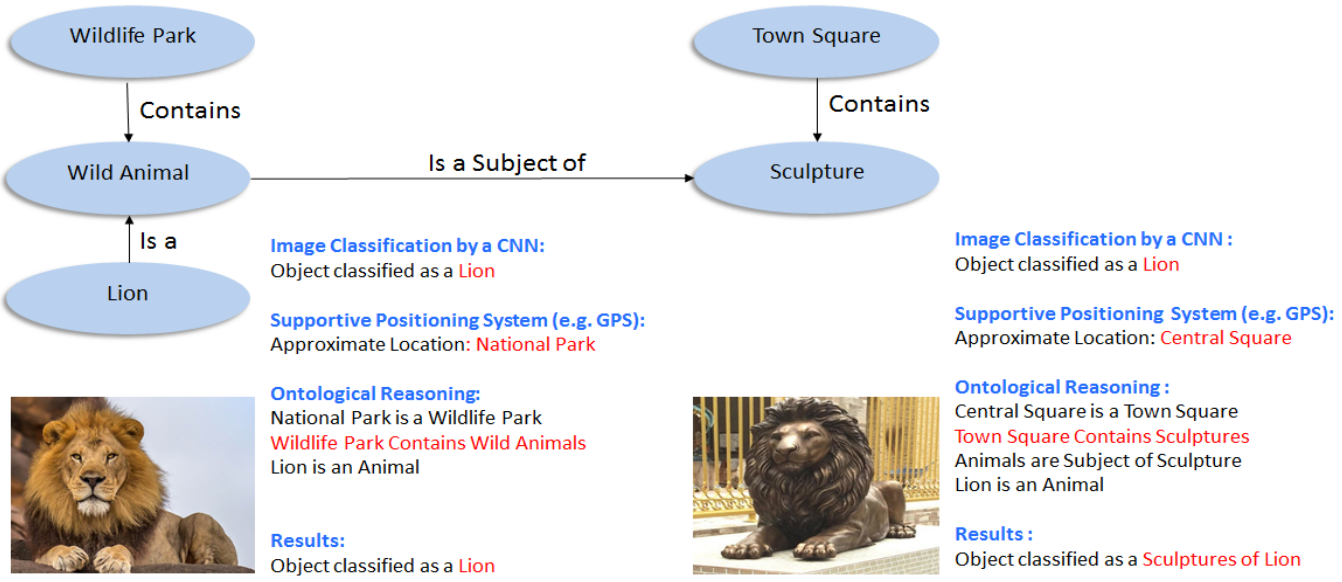

Figure 3. Employing background knowledge over image classification for image understanding.

Since an ontology represents abstract concepts and various relationships between such concepts, it can be used to model complex relationships between objects that are required for object recognition in applications such as computer vision and robotics. Gupta (2015) proposed an image classification approach using the context knowledge in an ontology over a Convolutional Neural Network. The CNN was trained by image features and then it was employed to determine weight values between the nodes represent features in the ontology.

Some studies in the literature used ontologies for semantic interpretations of objects. Breen et al. (2002) developed a Neural Network model for object recognition by employing a conceptual ontology on the sports domain that provides a semantic interpretation of images using associations between objects. Each sports terminology in the ontology contains a tag, a set of features, and weighted associations that can be used for interpreting sports images. In another research on object recognition (Tsai, 2012), an ontology is used for indoor and outdoor concepts. The results indicate that since such an ontology-based approach provides high-level semantic relations between concepts, it has advantages over traditional low-level feature extraction techniques for visual understanding.

Some other studies in the literature are focused on reasoning using ontological rules and restrictions. Burroughes and Gao (2016) proposed an ontology that contains rules and associations between objects in a topological map. These rules indicate that each object must be in certain locations on the map to be recognized. Another study on object recognition in an indoor environment (Schill et al., 2009) modeled the relationships between objects in a room using "part-of" and "has-a" relations to describe abstract structures of indoor objects. Such structures were used to recognize objects by their components.

In addition to simple relations, an ontology may contain probabilistic relations. In such cases, the model can be trained by analyzing the various rules, relations, and co-occurrence between concepts. These relations can be used for probabilistic reasoning.

Although a few studies on visual positioning took some advantages of semantic understanding; however, employing a comprehensive set of capabilities of a knowledge-based approach for describing concepts, presenting possibilities, defining rules, restrictions, and associations, can provide many advantages over traditional methods of visual positioning. In this regard, we propose a conceptual framework for visual positioning using various capabilities of ontologies.

3. A conceptual framework on employing a knowledge base over Neural Networks for visual positioning

To take advantage of context knowledge on visual positioning, we propose a conceptual framework to combine a CNN approach with an ontology as a knowledge base for positioning. As shown in Figure 4, in the data acquisition phase, pictures and the environmental data are acquired; then in the positioning phase, the user position is recognized by employing ontological rules, restrictions, and associations besides a two-level image classification approach using CNN. The framework is primarily designed for outdoor positioning by visually impaired people in urban areas where GPS signals are not available or when an exact position of the user is required; however, it also can be employed in other positioning applications in robotics, self-driving vehicles, etc.

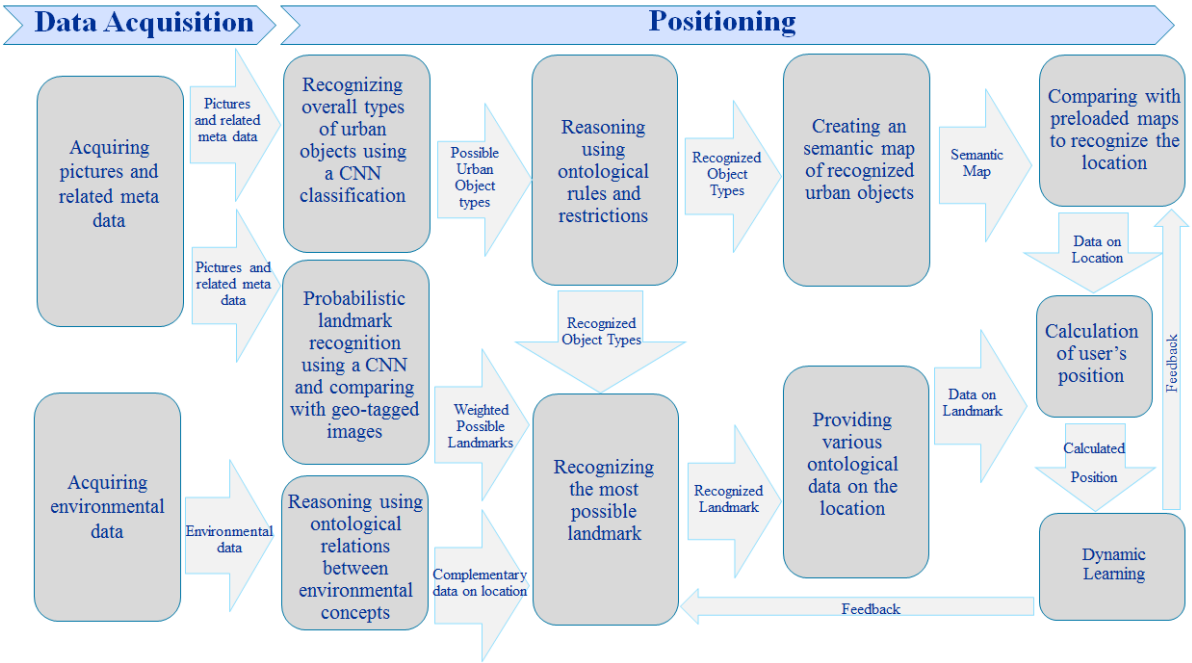

Figure 4. Proposed conceptual framework for visual positioning using ontology

In the first phase (data acquisition), a cellphone or any image capturing device can be used to acquire pictures. It is possible to use multiple pictures of a scene for positioning. These pictures can be assessed independently to crosscheck the results.

In addition to images, other environmental data such as sound and olfactory information can be acquired as complementary data. Such data and related ontological rules and associations provide more inputs for reasoning that empowers the positioning system.

Other than images and environmental data, it is also required to get the related meta-data of pictures. For example, an image has meta-data on direction and angle of the imaging device that is critical for an accurate visual positioning. Moreover, moments of the day and weather conditions determine which set of geo-tagged images are most suitable to compare with taken photos. A supportive positioning approach such as GPS or a recent location's data also can be helpful to provide an approximate location and narrow-down the search area of geo-tagged images.

In the second phase (positioning), the first level of classification by a CNN identifies the types of urban objects. At this level, abstract features of urban objects such as buildings, streets, etc. are used to train the model. After recognizing distinct types of urban objects, a semantic map of the location is provided by the system, which can be compared with a pre-loaded map of the region to determine the user's location. This can be done through five steps:

- Recognize type of urban objects in images;
- Produce a map of the environment using objects recognized in the images and by considering relations and rules between urban objects;
- A preloaded map of a region can be used to track consecutive positions of the user;
- Locations that are not likely to be the current position of the user can be filtered;
- Deduce the target location by comparing the semantic map with preloaded map and reasoning through ontological rules and restrictions;

The types of objects resulted by the first level of classification also can be used to narrow-down the target locations at the second level of CNN. The convolutional neural network in the second level is trained by geo-tagged images from landmarks and other urban objects. Then the model is used to classify input images and derive the position of the user by considering geo-tagged data included in the target dataset and also meta-data on the direction and angel of the imaging device.

Some studies (e.g., Kendall et al., 2015) indicate that training the model on both position and direction simultaneously leads to better results. It is also possible to use multiple geo-tagged images for a location in various moments of day and weather conditions to achieve the best results (e.g., Chen et al., 2016).

At this stage, environmental data also can be helpful to recognize possible locations. A few studies in the literature used environmental data for object recognition. Aguirre et al. (2013) used sound data for the detection of legs using noises of people's movements in an indoor environment. Monroy et al., (2019) also used environmental data besides visual information for detecting objects.

The second level of CNN needs huge datasets of geo-tagged images from different parts of urban areas to recognize locations. On the other hand, since a specific landmark or urban object is recognized by the system, more details will be available on the location. If coverage of geo-tagged images is not enough or no object related to those geo-tagged images is found in a taken photo, then the system should rely just on the first level of CNN for positioning. The overall architecture of this two-phase image classification approach is presented in Figure 5.

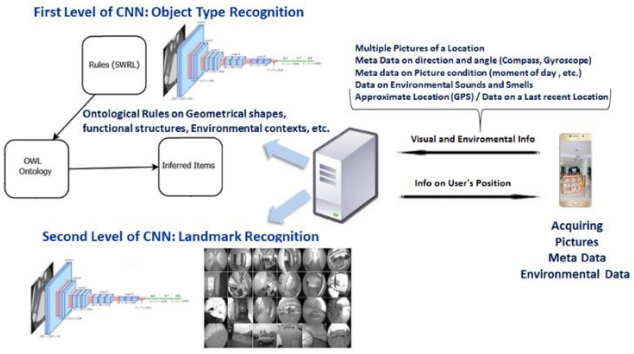

Figure 5. Overall architecture of the two-phase CNN-based classification

Since an ontology clearly describes concepts and also rules and relationships between such concepts, it can provide a powerful logical reasoning ability. This reasoning is an automated process, which is able to infer new axioms from current axioms in the knowledge base of an ontology. Any given relation and rule should not be contradicted with predefined rules and restrictions in the ontology. In addition, generalization and specialization associations and instance checking can be derived from relations such as "Part-of", "Is-a", "Has-a", etc.

The reasoning algorithm is also capable to use probabilistic relations in order to provide a list of the most possible objects or locations (Figure 6).

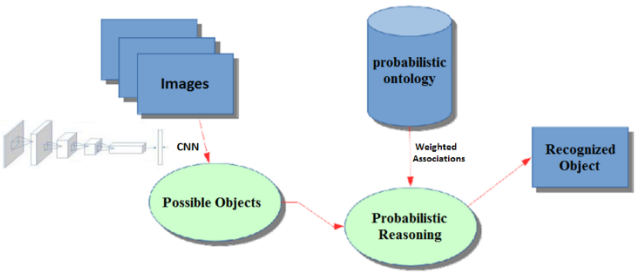

Figure 6. Reasoning using a probabilistic ontology

Reasoning can be performed based on the following steps:

- Defining a set of urban concepts in an ontology and creating associations between these concepts that realize any rules and restriction concerning urban objects;
- Employing verified results of an object recognition system to compute association weights between concepts and use them in a probabilistic ontology model;
- Getting all related rules and restrictions on associations between a given concept with others;
- Getting probabilistic weights of associations between a given concept with others;
- Considering rules and association for an urban concept to filter out object recognition false positives;
- Provide a list of most possible objects/locations using probabilistic weights of associations;

The last step of the framework is a dynamic knowledge interchange mechanism that can be used to give feedback on results and dynamically train the model. This could be done through a mechanism to receive users' captioning on images of new scenes and provide new information by users to actively retrain the model.

4. Applying the framework in an urban environment

To apply the framework in an urban environment, acquired images are feed to the first level of CNN that classifies the overall types of objects. In the following example (Figure 7) pictures of two different landmarks are recognized as towers.

Figure 7. First level of classification recognizes both objects as towers.

Then semantic maps of such locations are provided by the system, which can be compared with pre-loaded map of the region to determine the user's location. These maps can be completed through reasoning using ontological rules, associations and possibilities. Figure 8 presents an example of reasoning using ontological rules, restrictions and associations between urban objects. In our example, a tower is a building, and it should be above the street and also it is usually (weight: 0.8) taller than nearby houses.

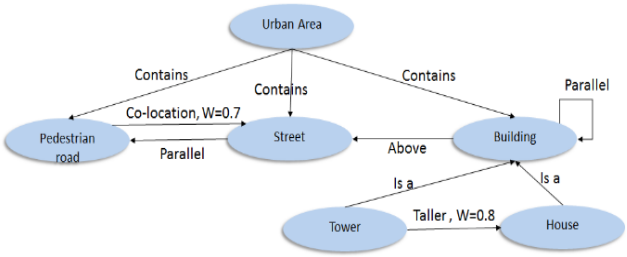

Figure 8. An example of rules and restriction in an urban ontology

The rules may also contain probabilistic weights that can be used for probabilistic reasoning. This can be useful for similar scenes where the object cannot be recognized with high accuracy. In such cases, a list of possible locations and also ontological rules and relations, besides environmental data, provide inputs for probabilistic reasoning to recognize the location with the highest possibility to be the current location of the user.

In parallel with the first CNN, the acquired images are also classified by the second level of CNN. Since the second level of CNN is trained by geo-tagged images from landmarks, the model classifies each object by similarities to the referenced images in the dataset. In our example, two pictures of towers are identified as different landmarks even though they have the same object type. The second level CNN is feed by the first level which recognizes overall types of objects; consequently, the search area is limited to landmarks with the type of tower in our example. If any reference image of the landmark exists in the dataset, this approach is capable to provide different ontological

information on the landmark such as relations with other urban objects and complementary information about the location. If the scene is not covered by the reference dataset, then the system uses the results of the first level classifier. Finally, the position of the user can be determined by considering the location of recognized landmarks and also meta-data on direction and angle of imaging devices.

In addition to two levels of CNNs, sound and olfactory meta-data of the location can be used to determine the most possible locations and filter out unlikely places. For example, the sound of motor vehicles implies a nearby street, or the smell of coffee may be related to a nearby coffee shop (Figure 9).

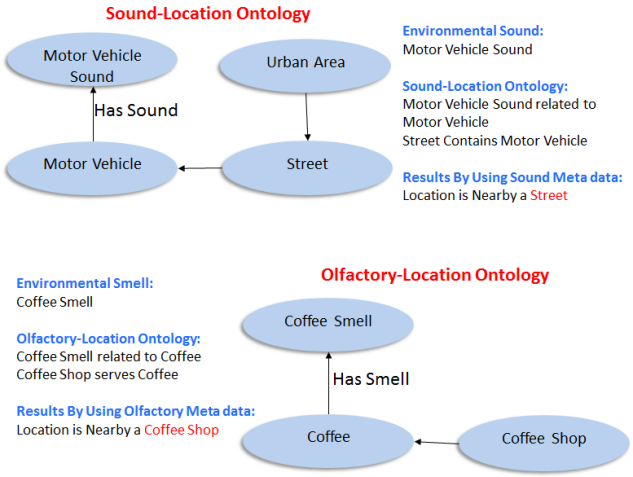

Figure 9. Relations of sound and olfactory data with their corresponding locations

The estimated position can used for dynamic training of the model. The taken images which do not exist in the geo-tagged dataset but are recognized through semantic map can be used for transfer learning to the model. Moreover, the positions which are not calculated properly can be assessed by system users to retrain the model by new inputs.

5. Conclusion and future research agenda

In recent years, positioning studies have had a greater focus on visual positioning which is empowered by new developments in image processing, and especially, deep learning techniques. Since pure deep learning techniques, lack context knowledge on locations, concepts, rules, and associations, such powerful image processing capabilities cannot simulate the human brain's ability to recognize places efficiently. Only a few studies in the existing literature, took advantages of a knowledge-based approach in visual positioning and even these studies concentrate on a few aspects of context knowledge.

In this paper, we presented a hybrid framework to employ ontologies over CNNs for integrating knowledge-based reasoning with neural networks. We also presented a practical example of visual positioning in an urban area that such a framework can be applicable. Comparing traditional methods of visual positioning which are focused on classifying images, our method provides a comprehensive set of capabilities including context knowledge, environmental data, object recognition, image classification, probabilistic reasoning, rules and associations to simulate the brain's functions for recognizing the position through a reasoning process.

Since the proposed framework can be employed in a variety of visual positioning applications such as robotics, self-driving vehicles, and positioning for visually impaired people, some additional steps may be appended to the framework by considering different requirements of such applications. Thereby, the next step would be tailoring the framework for different applications that requires more applied research in this field of study.

References

[1] Aguirre, E.; Garcia-Silvente, M.; Plata, J. Leg: Detection and Tracking for a Mobile Robot and Based on a Laser Device, Supervised Learning and Particle Filtering. In ROBOT2013: First Iberian Robotics Conference (2013).

[2] Aulinas, J., Petillot, Y., Salvi, J., Lladó, X.: The SLAM problem: a survey. In Proceedings of the 2008 conference on Artificial Intelligence Research and Development, 363–371 (2008).

[3] Breen C, Khan L, Kumar A.: Ontology-based image classification using neural networks. Proceedings of the ITCom 2002: The Convergence of Information Technologies and Communications (2002).

[4] Burroughes, G.,Gao, Y.: Ontology-based self-reconfiguring guidance, navigation, and control for planetary rovers. Journal of Aerospace Information Systems 13, 316–328 (2016).

[5] Chen, K., Wang, C., Wei, X.: Vision-Based Positioning for Internet-of-Vehicles, IEEE Transactions on Intelligent Transportation Systems 18 (2), pp. 364-376 (2016).

[6] Deretey, E., Ahmed, M. T., Marshall, J. A., Greenspan, M.: Visual indoor positioning with a single camera using PnP, International Conference on Indoor Positioning and Indoor Navigation (IPIN), pp. 1-9, doi:10.1109/IPIN.2015.7346756 (2015).

[7] Fan, J., Fang, L., Wu, J.: From Brain Science to Artificial Intelligence, Engineering, Volume 6, Issue 3, Pages 248-252 (2020).

[8] Goldman-Rakic P.S.: Cellular and circuit basis of working memory in prefrontal cortex of nonhuman primates, Prog Brain Res, 85, pp. 325-335 (1990).

[9] Gupta, U., Chaudhury, S.: Deep transfer learning with ontology for image classification, Proceedings of the Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, 1-4 (2015).

[10] Kendall, A., Grimes, Matthew, Cipolla, Roberto: Convolutional networks for real-time 6-DOF camera relocalization (2015).

[11] Kostavelis, I., Gasteratos, A.: Semantic mapping for mobile robotics tasks, Robotics and Autonomous Systems, Volume 66(C), pp 86–103 (2015).
[12] Li, B., Muñoz, JP., Rong, X., Chen, Q., Xiao J, Tian Y, Arditi A, Yousuf M: Vision-based Mobile Indoor Assistive Navigation Aid for Blind People. IEEE Trans Mob Computer, V18 (3):702-714 (2019).
[13] Li, B.; Wu, T.; Shuai1, S.; Zhang, L.; Chu, R.: Object Detection via Aspect Ratio and Context Aware Region-based Convolutional Networks, arXiv: 1612.00534v2 (2017).
[14] Liu, J., Feng-Ping, A.: Image Classification Algorithm Based on Deep Learning-Kernel Function Scientific Programming, 7607612 (2020).
[15] Monroy, J.; Ruiz-Sarmiento, J.; Moreno, F.; Galindo, C.; Gonzalez-Jimenez, J. Olfaction: Vision, and Semantics for Mobile Robots. Results of the IRO Project. Sensors 19, 3488 (2019).
[16] Rahman Su, Ullah S, Ullah S: A mobile camera based navigation system for visually impaired people. In: Proceedings of the 7th international conference on communications and broadband networking, pp 63–66 (2019).
[17] Schill, K.,Zetzsche, C., Hois, J.: A belief-based architecture for scene analysis: From sensorimotor features to knowledge and ontology, Fuzzy Sets and Systems, V 160(10), Pages 1507-1516, (2009).
[18] Studer, R., Benjamins, V., Fensel, R.: Knowledge engineering: principles and methods. Data & Knowledge Engineering. 25. 161-197 (1998).
[19] Tsai S F.: Toward ontological visual understanding. Dissertations & Theses - Gradworks (2012).