Robots with Commonsense: Improving Object Recognition through Size and Spatial Awareness

Agnese Chiatti[1], Enrico Motta[1] and Enrico Daga[1]

[1]*Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom*

Abstract

To effectively assist us with our daily tasks, service robots need object recognition methods that perform robustly in dynamic environments. Our prior work has shown that augmenting Deep Learning (DL) methods with knowledge-based reasoning can drastically improve the reliability of object recognition systems. This paper proposes a novel method to equip DL-based object recognition with the ability to reason on the typical size and spatial relations of objects. Experiments in a real-world robotic scenario show that the proposed hybrid architecture significantly outperforms DL-only solutions.

Keywords

commonsense reasoning, visual intelligence, hybrid intelligence, service robotics

1. Introduction

Robots can be of assistance in many scenarios where it is unsafe or impractical for us to intervene. For instance, they can help with waste disposal [1]; teleoperated patient care, particularly when social distance needs to be maintained [2]; search and rescue operations [3] and other tasks. However, operating in real-world scenarios is challenging because it requires to correctly interpret the diverse data collected through the robot's sensors [4]. Our approach to this *sensemaking* problem focuses on the modality of vision. That is, our focus is on equipping robots with high-performance *Visual Intelligence*, defined as "the ability to use their vision system, reasoning components and external knowledge sources to make sense of their environment" [5]. Consider HanS [6], the robot assistant that is currently undergoing development at the Knowledge Media Institute (KMi). HanS' job is to monitor the Lab in search of potential Health and Safety (H&S) threats: e.g., a paper cup left on top of a portable heater or a dangling cable in the middle of a corridor. As a necessary precondition to assessing correctly the risks posed by these situations, the robot first needs to reliably recognise the different objects involved, i.e., the paper cup and heater in question.

Currently, the *de facto* methodology for tackling object recognition tasks is to rely on Deep Learning (DL) methods. However, despite the breakthroughs that have been recently enabled

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2022 Spring Symposium on Machine Learning and Knowledge Engineering for Hybrid Intelligence (AAAI-MAKE 2022), Stanford University, Palo Alto, California, USA, March 21–23, 2022.

agnese.chiatti@open.ac.uk (A. Chiatti); enrico.motta@open.ac.uk (E. Motta); enrico.daga@open.ac.uk (E. Daga)

0000-0003-3594-731X (A. Chiatti); 0000-0003-0015-1592 (E. Motta); 0000-0002-3184-540 (E. Daga)

by DL methods, there are also significant limitations. In particular, as Cantwell Smith points out [7], state-of-the-art Artificial Intelligence (AI) technologies exhibit only a "reckoning" of the complete, deliberative thought processes that characterise human cognition. Hence, to shift from pattern recognition to higher-level cognition and sensemaking, robots need access to knowledge representations that are more comprehensive, transparent and explainable than those embedded in neural architectures [8, 9]. Consistently with this view, our previous work [5] has emphasised the key role played by commonsense knowledge in visual intelligence - see also [10, 11, 12, 13, 14]. In particular, as in the case of HanS, when robots are deployed in authentic problem-solving scenarios, it is important that they are both able to learn from their experiences and also able to asses the typicality and plausibility of a situation – e.g., that it is more likely to find a fire extinguisher, rather than a bottle, next to a fire extinguisher sign.

Specifically, in our prior work [5], we defined a set of required knowledge properties and reasoning capabilities that can enhance the Visual Intelligence of a robot. Among other things, this analysis pointed out that an awareness of both object sizes and spatial relations between objects has the potential to significantly improve the performance of object recognition systems. Hence, in [15], we developed a framework for representing object sizes, which can be used to enhance DL-based object recognition in the real-world scenario of autonomous H&S monitoring. Furthermore, in [16], we introduced a formal model of spatial representation, which fits the requirements of robot sensemaking. In the proposed representation, formally-defined spatial relations are mapped to a set of commonsense predicates, which are based on the types of prepositions that we use to describe and discuss space in natural language [17].

This paper builds on our prior work and introduces the following key contributions:

- we provide a methodology for extracting the commonsense spatial relations defined in [16] from large-scale multi-modal collections. In particular, we instantiate this approach in the case of the Visual Genome benchmark [18];
- we extend the work in [15] by automating the construction of a Knowledge Base (KB) about object sizes, capitalising on measurements gathered from ShapeNet [19] and Amazon;
- we discuss a robust evaluation of the resulting hybrid architecture, which integrates DL object recognition with size and spatial reasoning modules, in a real-world robotic scenario. In this context, we also test two meta-reasoning strategies, which provide alternative ways of combining the results of the reasoning modules introduced.

2. Related work

The work presented in this paper is situated at the intersection of different sub-fields of AI. First, we capitalise on the recent advances enabled by object recognition methods based on Deep Learning, and, in particular, on Deep Metric Learning methods, as further motivated in Section 2.1. Second, we contribute to the growing body of literature on developing hybrid intelligent systems, that can effectively leverage DL-based and knowledge-based components. Lastly, we contribute to the state-of-the-art in knowledge representation, in particular with respect to formalising and reasoning about size and spatial relations.

2.1. Object Recognition with Deep Metric Learning

Deep learning has enabled remarkable progress on several Computer Vision benchmarks [20, 21, 22]. Object recognition methods based on Deep Learning, however, suffer from certain drawbacks. These include - to name just a few: (i) the inability to learn from limited data; (ii) catastrophic forgetting, i.e., insufficient ability to retain the previously-learned information; (iii) the issue of disentanglement, i.e., limitations in discerning, recomposing, and reusing the separable constituents of observations; (iv) generalisation issues beyond the learned data distribution. The dependency of DL performance on the availability of large amounts of data, in particular, has inspired the development of *few-shot Metric Learning* methods. Deep Metric Learning (DML) is the task of learning a feature space where similar objects lie closely to one another and also further away from dissimilar objects. Specifically, in a few-shot learning scenario, only a few training examples are used. The typical DML setup consists of multiple pipelines based on Convolutional Neural Networks (CNN), yielding a so-called multi-branch network. Features extracted in each branch are compared through a similarity-based or through a distance-based function. In the case of Siamese Networks [23], two twin branches, fed with similar image examples, are updated with identical parameters. An extension of [23] are Triplet Networks [24], which are fed with both a positive and a negative example to contrast against the input image. Unlike traditional classification models based on DL, which are usually optimised over a pre-determined set of object categories, DML methods are targeted at comparing objects by similarity. Thus, the learned feature space can be more easily adapted as soon as novel object instances are observed. This characteristic of DML methods makes them particularly suitable for supporting robot sensemaking [25] – and indeed DML has been successfully applied in a number of robotic applications. In [26], a Triplet Network was used to recognise customers of a cafe in a human-robot collaboration scenario. In [27], a Siamese Network was used to derive optimal grasping poses for picking objects through a robotic arm. Zeng et al. [28] exploited DML to win the object stowing task at the latest Amazon Robotic Challenge. Thus, we will consider the methods presented in [28] as DL baselines. This choice also allows us to compare the results presented in this paper with our prior trials [15].

2.2. Hybrid Intelligent Systems

The limitations of DL methods have inspired the development of *hybrid methods*, which enhance DL by means of knowledge-based approaches. As described in [29], knowledge-based components can be integrated at four different levels of a Deep Neural Network: (i) in *pre-processing*, i.e., as input to the network; (ii) within the *intermediate layers*; (iii) as a part of the *architectural topology* or *optimisation function* or (iv) after applying DL, i.e., in *post-processing*. First, background knowledge can help to augment the training data. For instance, in [30], training examples were completed with Web-retrieved images of newly-encountered objects, to facilitate open-world object recognition. A second and increasingly popular trend in Computer Vision fuses visual data and symbolic knowledge within multi-modal embeddings [31, 32]. Encoder architectures have been used to combine visual features, extracted through a CNN, with a vectorised representation of the image context in a Knowledge Graph (KG) [32]. Similarly, in [31] object attributes have been encoded in multi-relational embeddings. However, these

methods suffer from the same disentanglement and explanation issues of traditional DL. Namely, the contribution of the different embedding features is difficult to pinpoint. More transparent than multi-modal embedding methods, approaches in the third group are aimed either at (i) modelling the reasoning process through the topology of a KG [33, 34], or at (ii) introducing modular knowledge-based components, that are trainable end-to-end [35]. For instance, [33] proposes to learn how to optimally traverse a KG via Deep Reinforcement Learning, to clarify the reasoning chains involved in Visual Question Answering. Another approach to hybrid reasoning is validating the DL predictions in post-processing [36, 37, 15]. Compared to hybrid methods based on end-to-end training, this class of methods offers the advantage of modularity, because it is agnostic to the specific DL architecture used. Moreover, it facilitates incremental knowledge updates by querying external KBs, as new object entities are observed. The work in [36] considered the object predictions generated via DL as a baseline to represent the spatial context of unknown objects. The system in [37] relied on expert knowledge to classify previously unseen classes, based on object parts detected via DL. Crucially, the post-hoc approach allows to decouple the DL predictions from the knowledge-based predictions and, thus, to more precisely assess how the different architectural modules contribute to the overall performance. This characteristic is an important precondition to identifying the strengths, weaknesses and complementarities of the DL-based and knowledge-based components. In [15] we proposed a system to validate DL predictions based on background knowledge about object sizes. Thanks to the modularity of the approach, we were able to conduct a detailed ablation study, to evaluate the effect of introducing size awareness over the baseline DL performance. In this paper, we build on this work, maintaining a post-hoc approach to hybrid reasoning.

2.3. Representing and reasoning about object sizes

Cognitive studies have suggested that the human brain maintains a *canonical* representation of the physical size of objects, which is functional to imagining and categorising them [38, 39, 40]. Inspired by the work of [38], [41] proposes to build a size graph, where nodes are modelled as log-normal distributions over Web-retrieved object dimensions. Similarly, [42] proposes to represent size as a statistical distribution of Web-retrieved measurements. [43] also collected physical size measurements from a combination of Freebase [44], Amazon and Ebay. Since the latter representations were aggregated from multiple measurements over large-scale databases, they are more likely to capture the *contour variance* among entities of the same class [40] – e.g., a short novel and a dictionary are both books, despite their size differences. However, these representations consider at most one feature contributing to size. Converging towards a single size feature is sufficient to identify broader groups of small and large objects, but hinders the classification of finer-grained object categories. For example, while the volume of a recycling bin may be comparable to the volume occupied by a coat stand, we would also need to know that the former is thicker than the latter, to discriminate between the two on the basis of size information. In [15], we proposed a representation of size information, where the cognitive features examined in [40] were condensed according to three coarse, class-level descriptors. These are: *surface area*, *thickness*, and *aspect ratio*. However, the resulting KB of standard object sizes, presented in [15], was manually curated. In this paper, we propose a method to extract these features automatically, from the size measurements provided with large-scale KBs.

2.4. Representing spatial relations between objects

The problem of building machine-treatable representations of space has been actively researched for decades, producing a plethora of formal spatial representations [45]. In robotics, much work has been devoted to grounding perceptual observations into higher-level symbolic representations [46, 47, 48, 9]. In this context, the *semantic map* representational model has been widely adopted [48, 9]. This is characterised as a map that, "in addition to spatial information about the environment", also contains "assignments of mapped features to entities of known classes" [46]. Attempts have been made at bridging the gap between semantic maps and formal spatial representations [36, 49, 50]. The authors of [36] focused on representing the closeness of objects in a semantic map through Ring Calculus. Kunze et al. [49] proposed to model spatial regions as point-like objects on the fixed 2D plane defined by a tabletop surface. Similarly, Deeken et al. [50] modelled a simplified scenario where the robot's viewpoint is aligned, at any point in time, with the coordinate system of the map, as well as with the orientation of the observed objects.

The latter representations exhibit two main shortcomings. First, they are unsuitable to model the case of a robot who is interpreting the environment whilst navigating it, because they are dependent on a fixed frame of reference, or coordinate system. Second, they define Qualitative Spatial Relations (QSR) - e.g., North of, West of, which are not aligned with natural language discourse, thus hindering Human-Robot Interaction. For instance, we want our robot to be capable of interpreting the instruction *bring me the cup that is on the left of the keyboard.* Moreover, mapping formal QSR to everyday linguistic predicates opens up the opportunity to reuse the spatial knowledge that is available within general-purpose Knowledge Bases, such as Visual Genome (VG) [18]. To address these shortcomings, we proposed the spatial framework described in [16], which consists of a set of First Order Logic (FOL) axioms that extend the definitions of [50], to account for varying viewpoints. Moreover, we also mapped the obtained set of relations to the everyday linguistic predicates examined in [17]. In this paper, we provide a methodology for (i) extracting the commonsense QSR described in [16] from large-scale KBs and (ii) integrating them in a concrete reasoning architecture.

3. Methodology

We propose to integrate an understanding of the typical size and spatial relations of objects with DL-based object recognition, through a post-hoc reasoning approach (Figure 1). In this setup, Deep Learning is applied first to derive a series of object predictions for each input image. Then, the observed size and spatial context of objects are considered, to infer the final object predictions. In Sections 3.1 and 3.2 we present a novel method to autonomously extract the relevant background knowledge from existing KBs. Then, we devote Section 3.3 to illustrating the different architectural modules in the proposed hybrid reasoning pipeline.

3.1. Representing the typical sizes of objects

One of the distinctive attributes of human Visual Intelligence is the ability to link the lower-level perceptual stimuli of the visual system with higher-level taxonomies of concepts [51, 40]. Thus, to characterise the size of objects, we first need to define a target object taxonomy. Another

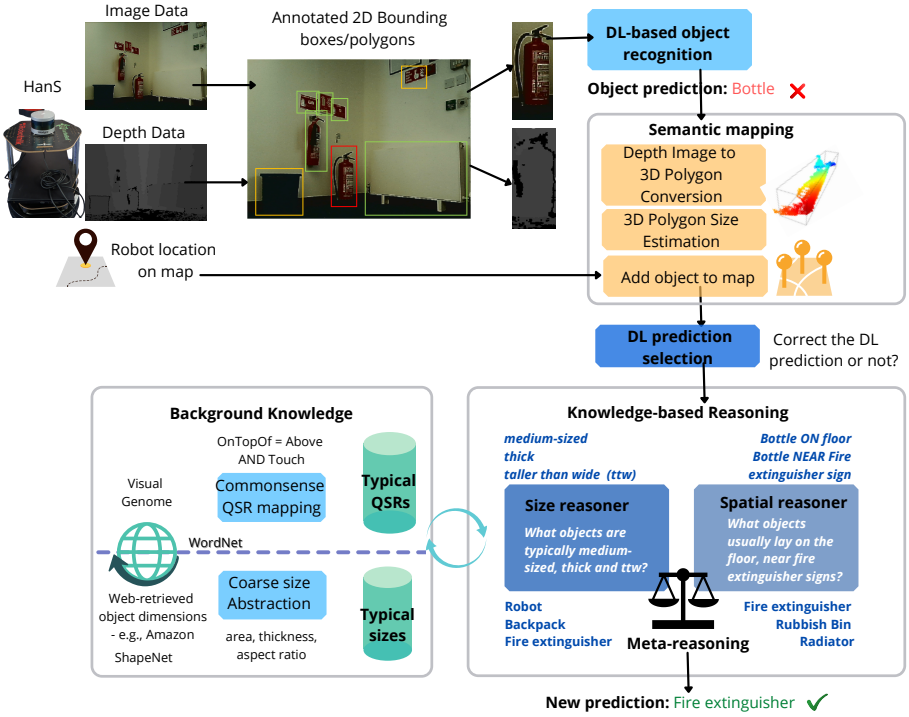

Figure 1: The proposed hybrid architecture for commonsense reasoning. Deep Learning is applied first to classify the objects in each image (top-left corner of the Figure). Then, objects are mapped in the 3D space based on the measured Depth data (in the "Semantic mapping" block). A subset of the predicted objects is passed to the "Knowledge-based Reasoning" module, which validates the predictions based on size and spatial knowledge.

requirement which can be derived from [40] is that the envisioned representation of size must be robust to *contour variance* - i.e., to variations in the shape and appearance of objects belonging to the same class. Specifically, in [40], object size is characterised in terms of *object extent*, *surface area*, *natural orientation* and *aspect ratio*. We propose to abstract these features from the raw size measurements available within general-purpose Knowledge Bases, as described in the following sections.

Taxonomy definition. Consistently with [15], we focus on 60 object classes that are commonly found at the Knowledge Media Institute, i.e., the environment where our robot assistant is currently deployed. To identify these categories, we examined the images collected by our robot during one of its scouting routines and identified the set of objects characterising the target environment. Objects which are too small to be detected through the robot camera (e.g., pens, light switches) were discarded. The remaining objects include furniture (e.g., chairs, desks), IT equipment (e.g., desktop PCs. monitors), H&S equipment (e.g., fire extinguishers, emergency exit signs) and other miscellaneous items. Specifically, we used domain-specific categories for H&S equipment and resorted to more generic class names otherwise. Intuitively,

the robot is expected to recognise that academic textbooks and notebooks are both types of books. At the same time, it is vital that the robot recognises the difference between emergency exit signs and fire extinguisher signs.

Handling contour variance. Based on how general is the object taxonomy defined at the previous step, an object class may include instances of varying size. Indeed, broad semantic concepts (e.g., chair) group objects of different shape and design (e.g., different chair models). To account for this intra-class variability, we collect repeated measurements from large-scale databases, as in [41, 42, 43]. Namely, for 31 out of the 60 target classes, the related object dimensions could be found on ShapeNet [19]. Additional object dimensions were scraped from Amazon. For the 14 remaining classes, measurements had to be collected manually, due to the paucity of comprehensive resources encoding size [41, 5]. To filter out noisy measurements, we applied Local Outlier Factor detection [52] across neighbour observation pairs in each class.

From instance-level absolute dimensions to class-level size features. Based on [40], the size of objects can be described in terms of area, orientation and extent, i.e., the ratio of the area of the object to its 2D bounding box. As such, the extent feature is affected by the so-called *framing effect* [38]: i.e., our perception of the size of an object is influenced by the size of the frame bounding the object. However, thanks to the availability of absolute object dimensions ($d1, d2, d3$), we can avoid to rely on the relative extent of objects within a frame to approximate their size. Nevertheless, to derive the area of an object from ($d1, d2, d3$), we need to make hypotheses about its orientation. For instance, we may consider a chest of drawers measuring 90x120x40 cm as 90 cm wide or as 40 cm wide, based on how it was placed. Because standard Web-retrieved measurements are represented as three object dimensions, there are only three possible area configurations that can be modelled: $d1{\cdot}d2$, $d2{\cdot}d3$, $d3{\cdot}d1$. As a result, the remaining dimension in each configuration indicates the *depth* (or *thickness*) of the object. To derive a summary descriptor of these features across instances of the same class, the area and depth values in each configuration are averaged class-wise. As the size variability across different classes can be significant (e.g., cupboards are significantly larger than drink cans), thickness and area values are also scaled with respect to their natural logarithm. This design choice also aligns with the cognitive findings of [38], which suggest that prototypical sizes are "proportional to the logarithm of the assumed size of the object".

Multi-feature abstraction. To group objects based on their size, we clustered the input class-level features on a 2D histogram, where the area is quantised over 5 bins (*extra-small*, *small*, *medium*, *large* or *extra-large*), and the thickness spans across 4 qualitative bins (*flat*, *thin*, *thick*, or *bulky*). An example of object sorting along these two dimensions is available at https://robots.kmi.open.ac.uk/img/size_repr.pdf. Because the resulting qualitative representation is expected to be human-understable and intuitive, we chose the most common format of Likert graphic rating scale for the object area, i.e., a 5-point scale with *medium* as a neutral category. For the thickness feature, we instead imposed a more polarised 4-point scale, as the excess of thickness, i.e., bulkiness, or lack thereof, i.e., flatness, are more distinctive object traits. In addition to these automatically-generated annotations, we also manually labelled classes *as flat* or *non-flat*. This caution allowed us to validate the thickness properties generated automatically. For example, if a certain object was labelled as strictly *flat* by the human oracle but categorised as *thick* through the histograms, the system would raise a conflict and correct

the latter annotation. Similarly, a second check condition was introduced to complete sparse annotations: e.g., if an object was annotated as both *medium* and *extra-large*, the *large* label would be automatically added too[1]. One last manual validation pass ensured that the relative sorting of objects is coherent. In addition to modelling different orientation configurations, we also considered that certain objects may be observed under *natural orientations* - e.g., chairs usually stand upright [5, 40]. Objects which are aligned with their natural orientation exhibit a distinctive *aspect ratio* (e.g., coat stands are typically taller than wide). Thus, we also labelled objects as *taller than wide* (*ttw*), *wider than tall* (*wtt*), or *equivalent*.

3.2. Representing the typical spatial relations of objects

Since robots can be mobile, and can observe objects from different viewpoints, in our earlier work [16], we devised a set of formally-defined Qualitative Spatial Relations (QSR) that support the modelling of different robot viewpoints. In our framework, QSR are expressed through first-order logic (FOL) axioms and 3D object regions are "contextualised", i.e., reconciled with the robot's current frame of reference. Moreover, our framework also provides a mapping from FOL axioms to the commonsense spatial predicates discussed in [17]. An excerpt of the proposed spatial relations and mapping is shown in Figure 2. We refrain from providing the complete definitions and FOL axioms that were presented in [16], for the sake of conciseness. Nonetheless, we ought to emphasise that the mapping from formal spatial relations to linguistic predicates is non-trivial, because spatial prepositions in English can be ambiguous. This issue is particularly pronounced in the case of the "on" preposition. For instance, we say that *a mug is on the counter* and that *a poster is on the wall*. Although "on" is used in both phrases, the spatial interpretation of the two sentences is very different. In the former case, the mug lays on top of the counter. In the latter case, the poster is affixed on the wall. To handle this polysemy, we identified three common uses of the "on" preposition, through the *OnTopOf*, *AffixedOn*, and *LeansOn* relations (Figure 2).

In the context of object recognition, we capitalise on this comprehensive knowledge model to reason on the typical spatial relations between objects. To this aim, we augment the QSR definitions defined in [16] with a measure of their typicality. Specifically, we propose to assess the typicality of a relation by measuring how frequently the relation occurs within a large-scale data collection. By relying on a large-scale set, the frequency of occurrence of a relation is less likely to reflect random effects or skews in the sample class distribution. In particular, Visual Genome (VG) [18] is one of the most comprehensive collections of spatial relations between objects [5]. It consists of 108,077 images, which are annotated with over 1.5 million object-object relationships. Spatial relations are prevalent in VG, as shown by its top ten most frequent predicates: on, has, in, wearing, wears, behind, next to, with, near, and in front of. Moreover, Visual Genome covers the majority of object categories in our taxonomy, i.e., 58 of the 60 classes. In the following, we focus on providing a definition of typicality, as well as a methodology to extract typical QSR from Visual Genome. The description of the module that supports reasoning about typical spatial relations in our architecture is instead given in Section 3.3.

[1]While this assumption is not always correct (e.g., cars can be extra-small, scaled models or extra-large, full-sized vehicles, even in the absence of medium-sized car models), it holds true for the class domains of this work. Hence, here we conservatively produced a larger set of annotations to avoid any size discrepancies.

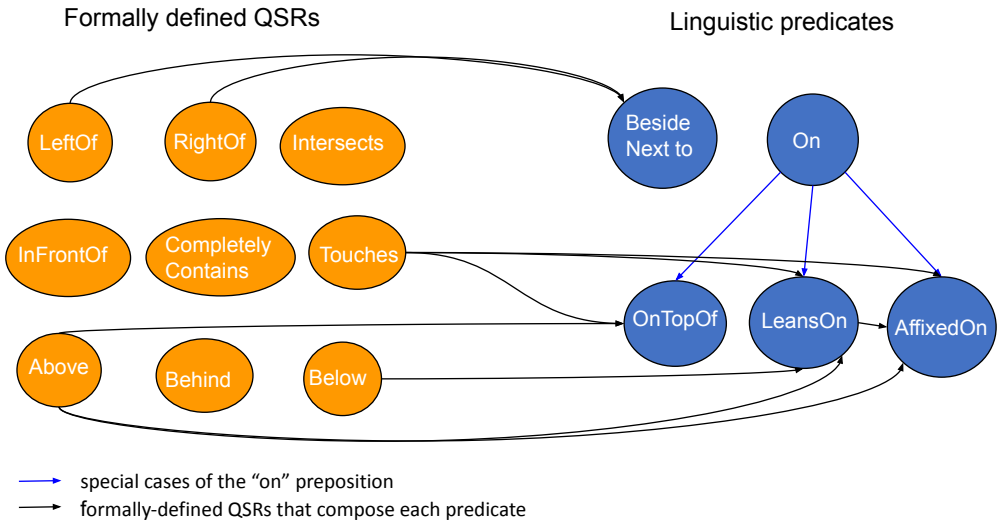

Figure 2: Sample mapping of formally-defined QSR to linguistic predicates. Predicates are linked to constituent QSR through black arrows. Blue arrows indicate special cases of the "on" preposition.

Typicality definition. The canonical structure of a spatial sentence [17, 16] is a triple where a figure object a (i.e., the grammatical subject) and a reference object b (i.e., the grammatical object) are linked through a spatial relation r. An example of canonical spatial relation is *coat stand is on the floor*. Let $N_{(a,r,b),C}$ be the number of times that a and b are in relation r within a data collection C. For instance, let us assume that the relation *coat stand is on the floor* occurs 100 times throughout VG. Moreover, let $N_{(a,r,\_),C}$ and $N_{(\_,r,b),C}$ indicate respectively: (i) the number of times that a is as the subject of predicate r, in C; (ii) the number of times that b is as the object of predicate r, in C. Namely, in VG, *coat stand* is the subject of *on* relations in 200 cases, and *floor* is the object of *on* relations in 500 cases. Then, the typicality of relation (a, r, b) in collection C can be defined as:

$$typicality_{(a,r,b),C} = \frac{N_{(a,r,b),C}}{N_{(a,r,\_),C} + N_{(\_,r,b),C} - N_{(a,r,b),C}}, \tag{1}$$

where $N_{(a,r,b),C} \leq N_{(a,r,\_),C}$ and $N_{(\_,r,b),C} \leq N_{(a,r,b),C}$. Namely, typicality is here defined as the *Jaccard similarity coefficient* between a and b. In our example, the *coat stand is on the floor* relation has a typicality score of $100/(200 + 500 - 100) \approx 0.17$. The more frequently two objects appear in relation r across the given collection, the more typical their relation is, i.e., the closer to 1 will be their Jaccard similarity score. Conversely, a relation which never appears throughout the collection yields a score of 0. From Equation 1, it follows that the *atypicality* of a relation can be expressed as the complement to 1, with respect to the chosen typicality metric, i.e., as the *Jaccard distance* between a and b, in terms of relation r.

Mapping VG predicates to Commonsense QSR. To derive a typicality score for the QSR defined in [16], these QSR first need to be mapped to the spatial predicates in Visual Genome. This process is complicated by the fact that spatial predicates in VG are often linked to incorrect aliases. For instance, the *above* and *on top of* predicates are treated as synonyms in VG.

Predicate keyword	Aliases	QSR
on top of		*OnTopOf*
against		*AffixedOn* $\vee$ *LeansOn*
beside	*next to*, *adjacent*, *on side of*	*Beside*
below	*under*	*Below*
above		*Above*
right, *left* *front*, *behind*		*RightOf*, *LeftOf* *InFrontOf*, *Behind*
near	*by*, *around*	*Near*

Table 1
Mapping of spatial keywords in Visual Genome, and their aliases, to the Commonsense QSR of [16].

However, these two predicates carry a very different spatial connotation: only the *on top of* expression implies that the two objects are in contact. To address this issue, we regrouped VG predicates based on key terms that could be directly mapped to the commonsense QSR in [16], as summarised in Table 1. We used keywords instead of key phrases (e.g., *front* instead of *in front of*), to account for the linguistic variability of VG predicates. For instance, the keyword "front" is shared by both the *in front of* and the *on the front of* predicates, which we want to map to the *InFrontOf* QSR. Here, each VG predicate is mapped to its nearest matching keyword. In the absence of exact matches, the matching sentence of maximum length is considered, so that, e.g., the relation "on the front of" is linked to the "front" preposition, as opposed to the "on" preposition. At this stage, we filter out i) relations that miss either the subject, the object or the predicate of the sentence, as well as ii) predicates expressed through the overly generic "of" and "to" prepositions.

The case of "on". To map occurrences of "on" to the correct spatial QSR, examining textual annotations is insufficient and further spatial inference on the input image is needed. To this aim, the 2D bounding boxes in each VG image are added to a spatial database, implemented in PostgreSQL [2]. The built-in spatial operators provided with the PostGIS engine[3] can then be used to conclude whether the two object regions involved in the "on" relation also overlap along the vertical direction. Only the occurrences of "on" which imply contact along the top of the reference object are mapped to the *OnTopOf* relation. The remaining occurrences are mapped to the "against" preposition, as instances of either the *AffixedOn* or *LeansOn* relations (Table 1).

Computing typicality scores. Given the definition of typicality introduced at Equation 1, each occurrence of a spatial relation found in VG increments the $N_{(a,r,b),C}$, $N_{(a,r,\_),C}$ and $N_{(\_,r,b),C}$ counts by one. Aliases in the same keyword group (Table 1) contribute to the same counts. For instance, hits of the "around" and "by" terms count as occurrences of the "near" relation. More broadly, all retrieved QSR, except for the "above" relation, contribute to the counts of the "near" relation. Indeed, object-object relations appearing in a natural scene generally imply proximity between the depicted objects. The "above" preposition, however, is an exception to this broader norm: e.g., *the sky is above*.

[2] https://www.postgresql.org/
[3] https://postgis.net/

3.3. Hybrid Architecture for Commonsense Reasoning

The proposed reasoning architecture (Figure 1), consists of the following components.

DL-based object recognition. This module classifies the objects depicted in each image. It receives as input an RGB image collected by the robot and a set of 2D object regions. Then, objects are classified through a Deep Neural Network that returns, for each input region, a set of ranked object predictions. In sum, this configuration is suitable for any algorithm which provides, for each detected object: (i) the 2D region bounding the object, (ii) a set of scored predictions, whether similarity-based or probability-based. Because the main aim of the experiments discussed in this paper is to compare the performance of hybrid solutions against the DML classification methods of [28], we rely on manually-annotated object regions, to control for potential errors propagated from the image segmentation steps.

Semantic mapping. The object predictions and associated depth data are fed to a second module, which maintains a *semantic map* of the robot's environment. In practice, by measuring the distance between the robot's position and the surfaces reached by the laser sensor, we can construct a *PointCloud*, i.e., a set of 3D geometrical points representing the input image. Then, the 2D object regions annotated on the RGB images are projected on the PointCloud, to derive 3D object regions and locate these regions on the geometrical map of the environment. These 3D regions are stored in a PostgreSQL spatial database and used to estimate the object dimensions. Lastly, the semantic map is completed by manually annotating the wall surfaces.

DL prediction selection. The next step is identifying which DL predictions need to be validated through knowledge-based reasoning. To this aim, the DL predictions are filtered by confidence. Specifically, because, in the considered baseline methods, object predictions are ordered by ascending L2 distance, we feed the knowledge-based reasoner only with those predictions whose top L2 distance, in the ranking, is greater than a threshold ϵ. Optimal values for the ϵ parameter are automatically estimated, through n-fold stratified cross validation. Specifically, a portion of the test set is held out, at each inference run, to derive the minimum L2 distance for each class. Minimum L2 values are averaged across classes, to infer the recommended ϵ.

Size reasoner. This module validates the DL predictions on the basis of the Knowledge Base of typical sizes constructed in Section 3.1. Hence, here the object sizes stored in the semantic map are quantised, so that they are expressed qualitatively, in terms of surface area, thickness and aspect ratio. The background size KB is queried to retrieve a set of object classes that are plausible with respect to the observed size. Then, the input DL ranking of object predictions is filtered so that only those classes deemed as valid in terms of size are retained. As a result, the output ranking of top-K predictions may include a different set of classes from those obtained through DL, but the original DL scores, i.e., the class order, is preserved.

Spatial reasoner. This reasoner considers the spatial relations of neighbour objects to correct the DL predictions. In particular, for each object instance that is selected for correction, the set of objects within a distance radius, R, is retrieved from the semantic map. We rely on PostGIS operators to compare neighbour object regions and derive their QSR, as described in [16]. The object neighbourhood is modelled as a directed graph, where nodes indicate object entities and edges encode spatial relations. As such, the efficacy of the spatial reasoner also depends on the quality of the neighbour object labels in the graph. When all object predictions

are generated automatically, DL errors can hinder the spatial reasoning results. For instance, consider the case of Figure 1, where a fire extinguisher was mistaken for a bottle. In principle, the proximity of a fire sign is a strong indication that the object is in fact a fire extinguisher. However, if the fire sign was mistaken, e.g., for a monitor, the same spatial cue would become misleading. For these reasons, in the experiments of Section 4, we control for the presence of DL-generated object labels. For each object prediction in the ranking and spatial relation in the graph, the typicality score is computed (Equation 1). To contain the computational cost of this step, typicality scores are computed only for the top-K predictions. Then, only the object predictions of non-null typicality score are retained and converted to atypicality scores, so that they can be combined with the L2 distances in the DL ranking, as follows. Atypicality scores are first averaged class-wise, across all observed QSR. Second, the class average is added to the L2 distance of the current prediction and re-sorted in ascending order. Hence, differently from the size reasoner, the set of predicted classes is always identical to the top-K DL ranking, whereas their ranked order may change. The aforementioned spatial reasoning steps are skipped in all cases where (i) no QSR are found for a given prediction, (ii) the DL predicted that the target region is a "person". Indeed, people can be found at various locations. Hence, the "person" class would always rank higher than other classes, invalidating the spatial reasoner results.

Meta-reasoning. Because the size reasoner and the spatial reasoner may recommend a different set of object classes, a rationale is needed to combine the outcomes of the two reasoners. In the experiments of this paper, we test two strategies for leveraging these reasoners. First, we apply the size and spatial reasoners in sequence, i.e., following a "waterfall" approach. Since the size reasoner only applies a filter to the input ranking, the order of execution of the two reasoners could be equivalently swapped. However, we also consider a second meta-reasoning scenario, which we refer to, in the following, as "parallel". In this scenario, the DL and size-validated rankings are fed to the spatial reasoner in parallel, to filter out the spatially-invalid predictions (see also Section 4.3).

4. Experiments

The experiments of this section are aimed at answering three main research questions. First, *are the performance benefits of integrating object recognition with size awareness, as measured in our prior work [15], preserved when the construction of the size knowledge base is automated?* Second, we interrogate on *the utility, in terms of object recognition performance, of introducing a second reasoning component, which considers the typical spatial relations between objects.* Lastly, we ask *what are the performance effects of combining both types of reasoners?*

As anticipated in Section 3.3, the performance of the spatial reasoner is dependent on the quality of the predictions generated for the nearby objects. Therefore, to control for error propagation from the DL-based to the knowledge-based steps while addressing our research questions, we set up two experiments. In **Experiment A**, objects which lie nearby each DL prediction are represented with ground truth labels. Namely, in the fire extinguisher example of Figure 1, the incorrect DL prediction, *bottle*, is associated with the spatial relation *near fire extinguisher sign*, irrespective of which class the DL algorithm predicted for the fire extinguisher sign. In **Experiment B**, no prior knowledge is available about the ground truth object categories. In

this realistic scenario, all object labels in the QSR graph are generated through DL.

The remainder of this section illustrates the dataset (Section 4.1), performance metrics (Section 4.2) and object recognition methods (Section 4.3) that were considered for evaluation. Lastly, the experimental results are presented and discussed in Section 4.4. The code supporting these experiments is available at https://github.com/kmi-robots/spatial-KB/tree/test.

4.1. The KMi RGB-D set

For consistency with our prior trials, we consider the same dataset introduced in [15]. In brief, images and depth data of the KMi Lab were collected through a Turtlebot mounting an Orbbec Astra Pro RGB-Depth (RGB-D) monocular camera. Specifically, 10 image examples per class were devoted to training the baseline models of [28] and to validating the model parameters, through an 80-20 training-validation split. Training examples were collected against a neutral background and cropped to reduce the marginal noise. Additionaly, test examples were sampled from one of our robot's monitoring routines. From the initial test sample, we annotated 1414 object regions, which were labelled with respect to the 60 classes of interest.

4.2. Evaluation metrics

In the experiments of this paper, performance is measured in terms of cross-class accuracy, precision, recall and F1 score of the top-1 object predictions in the ranking. Additionally, to account for the natural class imbalance in the KMi set [15], we also weigh these average metrics by class support. The quality of the top-5 predictions is evaluated in terms of mean Precision (P@5), mean normalised Discounted Cumulative Gain (nDCG@5) and hit ratio (HR), i.e., the number of times the correct prediction appeared among the top-5, divided by the total number of predictions. Evaluation metrics are averaged across 7 cross-validation splits. The value of n=7 was chosen pragmatically, to devote only a restrained portion of test examples to parameter tuning, i.e., 202 object regions out of 1414.

4.3. Ablation study

In the following, we contrast the performance of methods which are purely based on DL with variations of the proposed hybrid system. We start by evaluating the size and spatial reasoners individually. Then, we evaluate two meta-reasoning strategies that combine both reasoners.
N-net [28] is a Triplet Network based on three ResNet50 branches [53]. A feature space is learned so that exemplars of visually-similar objects lie closer, in terms of L2 distance, than dissimilar objects. This training objective is achieved by minimising the Triplet loss. At test time, objects are matched against their nearest neighbour in the learned feature space.

K-net [28] is identical to the N-net pipeline, except a second component is added to the Triplet loss during training. This component is based on the Softmax loss over M pre-defined classes. In this way, the training objective is modified to learn a feature space that can discriminate not only similar and dissimilar objects, but also a pre-defined subset of classes.

Hybrid (size only). In this variation of the workflow proposed in Section 3.3, only the size reasoner is considered. This method integrates knowledge of the object typical surface area,

thickness and aspect ratio, to validate the predictions generated by the DL algorithm. It resembles the "area+thin+AR" pipeline of [15], except here both the size knowledge construction and the estimation of the optimal confidence threshold, ϵ, are automated.

Hybrid (spatial only). In this second variation of the proposed architecture, only the spatial reasoner is considered. Specifically, the DL rankings selected for correction are modified based on: (i) the observed QSR, which are extracted from the input image and organised in a graph (Section 3.3), and (ii) the spatial typicality scores derived from Visual Genome (Section 3.2). Specifically, L2 distances in the DL ranking are combined with the Jaccard distances indicating the spatial atypicality of each class, yielding a modified ranking of predictions.

Hybrid (size + spatial, waterfall). In this implementation of the proposed hybrid reasoning framework, the DL, size-based and spatial reasoners are applied in sequence. First, the DL predictions are filtered based on the size reasoning outcomes. Then, the size-validated ranking is passed through the spatial reasoner. At this stage, the DL-based scores in the ranking are combined with the atypicality scores of the observed QSR.

Hybrid (size + spatial, parallel). Differently from the waterfall strategy, this method maintains the original DL predictions and the size-validated predictions as separate rankings. For instance, let us consider a case where the top-5 predictions in the DL ranking are *bottle*, *fire extinguisher*, *backpack*, *bin*, and *cup* - let this be ranking r_1. Once r_1 is fed to the size reasoner, only those predictions which are deemed as valid w.r.t. size are retained, yielding a second ranking, r_2: e.g., *fire extinguisher*, *backpack*. r_1 and r_2 are passed in parallel to the spatial reasoner. At this stage, the predictions of atypicality score 1 are filtered out of both rankings, yielding $\hat{r_1}$ - e.g., *fire extinguisher*, *bin* and $\hat{r_2}$ - e.g., *fire extinguisher*. Lastly, majority voting is used to derive the final predictions from $\hat{r_1}$ and $\hat{r_2}$. In our example, the *fire extinguisher* class would be selected as the final prediction, because it occurs twice. In the case of ties, only the size-validated predictions are considered.

4.4. Results and Discussion

Results are reported in Table 2. The DL baseline results are common to both experiments, A and B, because the DL module is independent from variations introduced in the reasoning modules. Similarly, the "size only" results are reported only once for both experiments, as the only difference between the two experiments pertains the spatial reasoner. The DL results reproduce the findings of [15]. Namely, K-net outperforms N-net, thanks to the optimisation over a pre-defined set of object classes. Therefore, in the following, we consider the predictions generated through K-net as input to the hybrid methods.

Importantly, the results achieved with the size reasoner match the top-1 accuracy and weighted R of [15] and outperform prior results on all remaining metrics. Thus, in relation to our first research question, although part of the knowledge base construction and the parameter tuning were automated, the baseline DL performance was enhanced by a significant margin. In particular, the unweighted and weighted F1 scores increased by 6% and 5%, compared to K-net. Interestingly, the precision of the size reasoner is higher than its recall. In the case of top-1 weighted metrics, this trend resembles the baseline K-net performance. Thus, a first explanation for this evidence is that the hybrid method maintains part of the prediction patterns of the

Table 2
Experiments A and B results on the KMi RGB-D test set.

		Top-1 unweighted			Top-1 weighted			Top-5 unweighted		
Method	Top-1 Acc.	P	R	F1	P	R	F1	P@5	nDCG@5	HR
N-net [28]	.45	.34	.40	.31	.62	.45	.47	.33	.36	.63
K-net [28]	.48	.39	.40	.34	.68	.48	.50	.38	.41	.65
Hybrid (size only)	.51	.47	.39	.40	.69	.51	.55	**.42**	.44	.68
Experiment A:										
Hybrid (spatial only)	.49	.41	**.43**	.37	.69	.49	.53	.38	.41	.65
Hybrid (waterfall)	.53	**.49**	.39	.41	**.71**	.53	.57	.39	.42	.68
Hybrid (parallel)	**.55**	.48	.42	**.42**	**.71**	**.55**	**.59**	**.42**	**.45**	**.69**
Experiment B:										
Hybrid (spatial only)	.48	.40	.41	.35	**.71**	.48	.52	.38	.41	.65
Hybrid (waterfall)	.52	.48	.39	.40	**.71**	.52	.56	.39	.42	.68
Hybrid (parallel)	.54	.48	.40	.41	**.71**	.54	.58	.41	.44	.68

underlying DL component. However, by definition, the size reasoner is "a more strict" classifier than the DL baseline: it classifies an object correctly when it observes that the object size is valid but also misses out a portion of true positives when the measured size is found to be non-valid. Wrong size judgements could be equally caused by (i) errors in the estimation of the object size, by (ii) biases in the distribution of object sizes in the raw data, or by (ii) the methodological choice to adopt discrete size categories. For instance, if an object lies at the frontier between small and medium-sized objects, it may be wrongly categorised due to relying on clear-cut thresholds.

Experiment A. As shown in Table 2, the spatial reasoner also contributes to a performance improvement over the DL baselines. The F1 scores achieved in the "spatial only" case outperformed K-net by 3%. On average, the improvement is more modest than in the case of the size reasoner. These results align with the hypothesis we laid in [5]: that spatial reasoning capabilities are relatively less impactful towards object categorisation than size reasoning. Interestingly, the ranking quality metrics in the "spatial only" scenario are equivalent to the K-net case. This result can be explained by considering that the spatial reasoner only re-ranks the DL predictions, without discarding any of the originally predicted classes. As such, it may overlook potential candidates that lie outside the top-K positions, compared to the size reasoner. Nonetheless, this characteristic also ensures more balanced unweighted P and R scores than the "size only" case. Combining both reasoners leads to a higher, or at least comparable performance than that achieved through the individual reasoners, across all the top-1 metrics. The DL-only baseline was surpassed by a significant margin. The "parallel" strategy, overall, led to the highest margin of improvement: the K-net accuracy improved by 7%; the unweighted metrics improved by 9% (P), 2% (R) and 8% (F1); the weighted metrics increased by 3% (P), 7% (R), 9% (F1). All the top-5 quality metrics also increased by 4%. Moreover, the unweighted P and R are more balanced in the "parallel" than in the "waterfall" scenario. Because, in the "waterfall" case, the size and spatial reasoner are applied in sequence, the same performance trend of the "size only" case

emerges. Conversely, in the "parallel" case, when the DL and size-validated predictions are leveraged through the spatial reasoner, the higher precision ensured by the size reasoner is coupled with the higher recall trend of the spatial reasoner, yielding the highest F1 results across the tested hybrid methods.

Experiment B. In this realistic scenario, the object labels for the observed QSR are generated automatically, through DL. Therefore, as expected, the performances of the "spatial only", "waterfall" and "parallel" methods were lower than in Experiment A. Despite this slight performance degrade, the hybrid methods which integrate both size and spatial awareness still ensured a significant performance improvement over the K-net baseline. Moreover, Experiment B confirmed the performance patterns already observed in Experiment A. First, the "parallel" method achieved the highest performance across the majority of evaluation metrics, with the exception of the unweighted R and P@5, which are 1% lower than other hybrid methods. Second, as in Experiment A, the performance gap between the precision and recall scores is less pronounced in the "parallel" than in the "waterfall" scenario. This evidence suggests that the "parallel" meta-reasoning strategy was the most effective one in terms of leveraging the advantages and shortcomings of its constituent reasoners.

5. Conclusion

We have presented a novel hybrid system that enhances the object recognition performance of a robot, by introducing awareness of the typical size and spatial relations of objects. Because the knowledge representations proposed in this work are designed for capturing the typicality of size and spatial properties, they relate to the broader objective of equipping intelligent systems with commonsense. We showed how commonsense knowledge can be autonomously extracted from general-purpose Knowledge Bases, namely: (i) ShapeNet and Amazon in the case of size, (ii) Visual Genome, in the case of spatial relations.

Furthermore, we have demonstrated how the extracted commonsense knowledge can significantly augment the performance of object recognition systems which are purely based on DL, in the real-world scenario of autonomous Health and Safety monitoring. Specifically, the most robust results were achieved when both types of knowledge-based reasoners were leveraged. The significant performance improvement achieved over DL is the key take-away message from this study. Since we proposed a modular solution, the absolute performance figures that have been presented only provide a reference. Naturally, other DL solutions could be integrated in our system, to improve the quality of the initial predictions. In our future work, we are also interested in assessing whether a specific type of reasoner was most helpful to recognise certain subsets of classes, especially with respect to categories that are specific to Health & Safety. To this aim, we will evaluate the proposed system on H&S decision-making tasks, such as, e.g., verifying the correct placement of fire extinguishers. Another extension of this work concerns the problem of reconciling successive observations of the same object entity. In this paper we considered a basic model of semantic map, where reasoning is performed on a frame-by-frame basis. However, tracking objects across frames and updating the mapped observations incrementally would further improve the robot's object recognition abilities.

References

[1] M. J. C. Samonte, S. H. Baloloy, C. K. J. Datinguinoo, e-tapon: Solar-powered smart bin with path-based robotic garbage collector, in: 2021 IEEE 8th International Conference on Industrial Engineering and Applications (ICIEA), IEEE, 2021, pp. 181–185.

[2] G. Yang, H. Lv, Z. Zhang, L. Yang, J. Deng, S. You, J. Du, H. Yang, Keep healthcare workers safe: application of teleoperated robot in isolation ward for covid-19 prevention and control, Chinese Journal of Mechanical Engineering 33 (2020) 1–4.

[3] B. R. Le Comte, G. S. Gupta, M. T. Chew, Distributed sensors for hazard detection in an urban search and rescue operation, in: 2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings, IEEE, 2012, pp. 2385–2390.

[4] M. B. Alatise, G. P. Hancke, A review on challenges of autonomous mobile robot and sensor fusion methods, IEEE Access 8 (2020) 39830–39846.

[5] A. Chiatti, E. Motta, E. Daga, Towards a Framework for Visual Intelligence in Service Robotics: Epistemic Requirements and Gap Analysis, in: Proceedings of the 17th KR2020 Conference, IJCAI, 2020, pp. 905–916. doi:10.24963/kr.2020/93.

[6] E. Bastianelli, G. Bardaro, I. Tiddi, E. Motta, Meet hans, the heath & safety autonomous inspector., in: International Semantic Web Conference - Demo track, 2018.

[7] B. Cantwell Smith, The Promise of Artificial Intelligence: Reckoning and Judgement, The MIT Press, Cambridge, 2019.

[8] J. Pearl, Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution, in: Proceedings of the ACM WSDM Conference, WSDM '18, ACM, New York, NY, USA, 2018, p. 3. doi:10.1145/3159652.3176182.

[9] D. Paulius, Y. Sun, A survey of knowledge representation in service robotics, Robotics and Autonomous Systems 118 (2019) 13–30.

[10] E. Davis, G. Marcus, Commonsense reasoning and commonsense knowledge in artificial intelligence, Communications of the ACM 58 (2015) 92–103. URL: https://doi.org/10.1145/2701413. doi:10.1145/2701413.

[11] H. Levesque, Common Sense, the Turing Test, and the Quest for Real AI | The MIT Press, The MIT Press, 2017. URL: https://mitpress.mit.edu/books/common-sense-turing-test-and-quest-real-ai.

[12] P. J. Hayes, The Second Naive Physics Manifesto, Formal theories of the common sense world (1988). URL: https://ci.nii.ac.jp/naid/10022004548/, publisher: Ablex Publishing Corporation.

[13] J. McCarthy, et al., Programs with common sense, RLE and MIT computation center, 1960.

[14] A. Newell, The knowledge level, Artificial intelligence 18 (1982) 87–127.

[15] A. Chiatti, E. Motta, E. Daga, G. Bardaro, Fit to measure: Reasoning about sizes for robust object recognition, in: To appear in proceedings of the AAAI2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021), 2021.

[16] A. Chiatti, G. Bardaro, E. Motta, E. Daga, Commonsense spatial reasoning for visually intelligent agents, arXiv preprint arXiv:2104.00387 (2021).

[17] Barbara Landau, R. Jackendoff, "What" and "where" in spatial language and spatial cognition, Behavioral and Brain Sciences 16 (1993) 217–265.

[18] R. Krishna, Y. Zhu, O. Groth, J. Johnson, et al., Visual genome: Connecting language and

vision using crowdsourced dense image annotations, International journal of computer vision 123 (2017) 32–73.

[19] M. Savva, A. X. Chang, P. Hanrahan, Semantically-enriched 3D models for common-sense knowledge, in: 2015 IEEE CVPRW Conference, 2015, pp. 24–31. doi:10.1109/CVPRW.2015.7301289, iSSN: 2160-7516.

[20] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, M. Pietikäinen, Deep Learning for Generic Object Detection: A Survey, International Journal of Computer Vision 128 (2020) 261–318. doi:10.1007/s11263-019-01247-4.

[21] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, Communications of the ACM 60 (2017) 84–90. URL: https://doi.org/10.1145/3065386. doi:10.1145/3065386.

[22] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. URL: https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

[23] G. Koch, R. Zemel, R. Salakhutdinov, Siamese neural networks for one-shot image recognition, in: ICML deep learning workshop, volume 2, Lille, 2015.

[24] E. Hoffer, N. Ailon, Deep Metric Learning Using Triplet Network, in: A. Feragen, M. Pelillo, M. Loog (Eds.), Similarity-Based Pattern Recognition, Lecture Notes in Computer Science, Springer, Cham, 2015, pp. 84–92. doi:10.1007/978-3-319-24261-3_7.

[25] B. J. Meyer, T. Drummond, The importance of metric learning for robotic vision: Open set recognition and active learning, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 2924–2931.

[26] G. Sawadwuthikul, T. Tothong, T. Lodkaew, P. Soisudarat, S. Nutanong, P. Manoonpong, N. Dilokthanakul, Visual goal human-robot communication framework with few-shot learning: a case study in robot waiter system, IEEE Transactions on Industrial Informatics (2021).

[27] K. Suzuki, Y. Yokota, Y. Kanazawa, T. Takebayashi, Online self-supervised learning for object picking: detecting optimum grasping position using a metric learning approach, in: 2020 IEEE/SICE International Symposium on System Integration (SII), IEEE, 2020, pp. 205–212.

[28] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, et al., Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching, in: 2018 IEEE ICRA Conference, IEEE, 2018, pp. 1–8.

[29] S. Aditya, Y. Yang, C. Baral, Integrating Knowledge and Reasoning in Image Understanding, in: Proceedings of IJCAI 2019, 2019, pp. 6252–6259.

[30] M. Mancini, H. Karaoguz, E. Ricci, P. Jensfelt, B. Caputo, Knowledge is Never Enough: Towards Web Aided Deep Open World Recognition, in: IEEE ICRA Conference, 2019, p. 9543. doi:10.1109/ICRA.2019.8793803.

[31] A. Daruna, W. Liu, Z. Kira, S. Chetnova, Robocse: Robot common sense embedding, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 9777–9783.

[32] G. Castellano, G. Sansaro, G. Vessio, Integrating contextual knowledge to visual features

for fine art classification, in: Workshop on Deep Learning for Knowledge Graphs (DL4KG). The International Semantic Web Conference (ISWC)., 2021.
[33] R. Koner, H. Li, M. Hildebrandt, D. Das, V. Tresp, S. Günnemann, Graphhopper: Multi-hop scene graph reasoning for visual question answering, in: International Semantic Web Conference, Springer, 2021, pp. 111–127.
[34] K. Marino, R. Salakhutdinov, A. Gupta, The More You Know: Using Knowledge Graphs for Image Classification, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 20–28. doi:10.1109/CVPR.2017.10, iSSN: 1063-6919.
[35] E. van Krieken, E. Acar, F. van Harmelen, Analyzing Differentiable Fuzzy Implications, in: Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning, 2020, pp. 893–903. URL: https://doi.org/10.24963/kr.2020/92. doi:10.24963/kr.2020/92.
[36] J. Young, L. Kunze, V. Basile, E. Cabrio, N. Hawes, B. Caputo, Semantic web-mining and deep vision for lifelong object discovery, in: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2017, pp. 2774–2779.
[37] G. J. Burghouts, F. Hillerström, Zero-detect objects without training examples by knowing their parts., in: AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering, 2021.
[38] T. Konkle, A. Oliva, Canonical visual size for real-world objects, Journal of Experimental Psychology: Human Perception and Performance 37 (2011) 23–37. doi:10.1037/a0020413.
[39] J. B. Julian, J. Ryan, R. A. Epstein, Coding of object size and object category in human visual cortex, Cerebral Cortex 27 (2017) 3095–3109.
[40] B. Long, T. Konkle, M. A. Cohen, G. A. Alvarez, Mid-level perceptual features distinguish objects of different real-world sizes., Journal of Experimental Psychology: General 145 (2016) 95.
[41] H. Bagherinezhad, H. Hajishirzi, Y. Choi, A. Farhadi, Are elephants bigger than butterflies? reasoning about sizes of objects, in: Proceedings of the Thirtieth AAAI Conference, AAAI Press, 2016, pp. 3449–3456.
[42] Y. Elazar, A. Mahabal, D. Ramachandran, T. Bedrax-Weiss, D. Roth, How Large Are Lions? Inducing Distributions over Quantitative Attributes, in: Proceedings of the 57th Annual Meeting of the ACL, Association for Computational Linguistics, 2019, pp. 3973–3983.
[43] Y. Zhu, A. Fathi, L. Fei-Fei, Reasoning about Object Affordances in a Knowledge Base Representation, in: Computer Vision – ECCV 2014, volume 8690, Springer, Cham, 2014, pp. 408–424. doi:10.1007/978-3-319-10605-2_27, series Title: Lecture Notes in Computer Science.
[44] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: Proceedings of the SIGMOD 2008, 2008, pp. 1247–1250.
[45] A. G. Cohn, J. Renz, Qualitative Spatial Representation and Reasoning, in: F. van Harmelen, V. Lifschitz, B. Porter (Eds.), Foundations of Artificial Intelligence, volume 3 of *Handbook of Knowledge Representation*, Elsevier, 2008, pp. 551–596.
[46] A. Nüchter, J. Hertzberg, Towards semantic maps for mobile robots, Robotics and Autonomous Systems 56 (2008) 915–926.

[47] S. Coradeschi, A. Saffiotti, An introduction to the anchoring problem, Robotics and autonomous systems 43 (2003) 85–96.
[48] I. Kostavelis, A. Gasteratos, Semantic mapping for mobile robotics tasks: A survey, Robotics and Autonomous Systems 66 (2015) 86–103.
[49] L. Kunze, C. Burbridge, M. Alberti, A. Thippur, J. Folkesson, P. Jensfelt, N. Hawes, Combining top-down spatial reasoning and bottom-up object class recognition for scene understanding, in: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2014, pp. 2910–2915.
[50] H. Deeken, T. Wiemann, J. Hertzberg, Grounding semantic maps in spatial databases, Robotics and Autonomous Systems 105 (2018) 146–165.
[51] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building machines that learn and think like people, Behavioral and Brain Sciences 40 (2017).
[52] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, Lof: identifying density-based local outliers, in: Proceedings of SIGMOD 2000, 2000, pp. 93–104.
[53] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.