# Fit to Measure: Reasoning About Sizes for Robust Object Recognition

Agnese Chiatti[a], Enrico Motta[a], Enrico Daga[a] and Gianluca Bardaro[a]

[a]*Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom*

**Abstract**

Service robots can help with many of our daily tasks, especially in those cases where it is inconvenient or unsafe for us to intervene – e.g., under extreme weather conditions or when social distance needs to be maintained. However, before we can successfully delegate complex tasks to robots, we need to enhance their ability to make sense of dynamic, real-world environments. In this context, the first prerequisite to improving the Visual Intelligence of a robot is building robust and reliable object recognition systems. While object recognition solutions are traditionally based on Machine Learning, augmenting them with knowledge-based reasoners has been shown to improve their performance. In particular, based on our prior work on identifying the epistemic requirements of Visual Intelligence, we hypothesise that knowledge of the typical size of objects can significantly improve the accuracy of an object recognition system. To verify this hypothesis, in this paper we present an approach to integrating knowledge about object sizes in a ML-based architecture. Our experiments in a real-world robotic scenario show that this hybrid approach ensures a significant performance increase over state-of-the-art Machine Learning methods.

**Keywords**

object recognition, service robotics, hybrid AI, reasoning about sizes, cognitive systems

## 1. Introduction

With the fast-paced advancement of the Artificial Intelligence (AI) and Robotics fields, there is an increasing potential to resort to *service robots* (or *robot assistants*) to help with daily tasks. Service robots can take on many roles. They can operate as patient carers [1], Health and Safety monitors [2], museum or tour guides [3], to name a few. However, succeeding in the real world is a challenge because it requires robots to make sense of the high-volume and diverse data coming through their perceptual sensors. Although different sensory modalities contribute to the robot's *sensemaking* abilities (e.g., touch, sound), in this work, we focus on the modality of vision. From this entry point, the problem then becomes one of enabling robots to correctly interpret the stimuli of their vision system, with the support of background knowledge sources, a capability also known as *Visual Intelligence* [4]. The first prerequisite to Visual Intelligence is the ability to robustly recognise the different objects occupying the robot's environment (*object*

*In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021) - Stanford University, Palo Alto, California, USA, March 22-24, 2021.*

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

0000-0003-3594-731X (A. Chiatti); 0000-0003-0015-1592 (E. Motta); 0000-0002-3184-540 (E. Daga); 0000-0002-6695-0012 (G. Bardaro)



*recognition*). Let us consider the case of HanS, the Health and Safety robot inspector currently under development at the Knowledge Media Institute (KMi). HanS is expected to monitor the Lab in search of potentially dangerous situations, such as fire hazards. Imagine that Hans was observing a flammable object (e.g., a paper cup) left on top of a portable heater. To conclude that it is in the presence of a potential fire hazard, the robot would first need to recognise that the cup and the heater are there.

Currently, the most common approach to tackling object recognition tasks is applying methods which are based on Machine Learning (ML). In particular, the state-of-the-art performance is defined by the latest approaches based on Deep Learning (DL) [5, 6]. Despite their popularity, these methods have received many critiques due to their brittleness and lack of transparency [7, 8, 9]. To compensate for these limitations, a more recent trend among AI researchers has been to combine ML with knowledge-based reasoning, thus adopting a *hybrid approach* [10, 11]. A question remains, however, on what type of knowledge resources and reasoning capabilities should be leveraged within this new class of hybrid methods [12].

In [4], we identified a set of *epistemic requirements*, i.e., a set of capabilities and knowledge properties, required for service robots to exhibit Visual Intelligence. We then mapped the identified requirements to the types of classification errors emerging from one of HanS' scouting rounds, where we relied solely on Machine Learning to recognise the objects. This error analysis highlighted that, in 74% of the cases, a more accurate object classification could in principle have been achieved if the relative size of objects was considered for their categorisation. This view is also supported by studies of human visual cognition, which suggest that our priors about object sizes are crucial to their categorisation [13, 14]. For instance, back to HanS' case, the paper cup could be mistaken for a rubbish bin, due to its shape. However, rubbish bins are typically larger than cups. With this awareness, HanS would be able to rule out object categories, which, albeit being visual similarity to the correct class, are *implausible* from the standpoint of size. These elements of *typicality* and *plausible reasoning* [15] link size reasoning to the broader objective of developing AI systems which exhibit *common sense* [16], and *intuitive Physics* reasoning abilities [17, 18].

On a more practical level, knowledge representations which encode object sizes have already been applied effectively to answering questions posed in natural language, such as "is A larger than B?" [19, 20]. However, despite this body of theoretical and empirical evidence, the role of size in object recognition has received little attention in the field of Computer Vision. In this paper, we investigate the performance effects of augmenting a ML-based object recognition with a reasoner which accounts for the size of the observed objects. In particular, we propose: • a hybrid method to validate ML-based predictions based on the typical size of objects; • a novel representation for size, which categorises objects based on their front surface area, depth and aspect ratio.

## 2. Related Work

State-of-the-art object recognition methods rely heavily on Machine Learning, as further discussed in Section 2.1. Because of the limitations of ML methods, hybrid methods, which combine ML with background knowledge and knowledge-based reasoning, have been recently

proposed (Section 2.2). In particular, our hypothesis is that awareness of the object size has the potential to drastically improve the performance of hybrid object recognition methods [4]. Therefore, in Section 2.3, we discuss the existing approaches to representing the size of objects.

## 2.1. Machine Learning for Object Recognition

The impressive performance exhibited by object recognition methods based on DL has led to significant advances on several Computer Vision benchmarks [5, 21, 22]. Deep Neural Networks (NNs), however, come with their limitations. These models (i) are notoriously data-hungry, i.e., require thousands of annotated training examples to learn from, (ii) learn classification tasks offline, i.e., assume to operate in a closed world [23], and (iii) learn representational patterns automatically, by iterating over a raw input set [6]. The latter trait can drastically reduce the start-up costs of feature engineering. However, it also complicates tasks such as explaining results and integrating explicit knowledge statements in the pipeline [7, 9].

The issue of robust learning from minimal training examples has inspired the development of few-shot metric learning methods. *Metric learning* is the task of learning an embedding (or feature vector) space, where similar objects are mapped closer to one another than dissimilar objects. In this setup, even objects unseen at training time can be categorised, by matching the learned representations against a support (reference) image set. In particular, in a *few-shot scenario*, the number of training examples and support images is kept to a minimum. Deep metric learning has been applied successfully to object recognition tasks [24, 25, 26], even in real-world, robotic scenarios [27]. Koch and colleagues [24] proposed to train two identical Convolutional Neural Networks (CNN) fed with images to be matched by similarity. This twin architecture is also known as Siamese Network. An extension of the Siamese architecture is the Triplet Network [25, 26], where the input data are fed as triplets including: (i) one image depicting a certain object class (i.e., *anchor*), (ii) a *positive example* of the same object, and (ii) a *negative example*, depicting a different object. The winning team for the object stowing task at the latest Amazon Robotic Challenge capitalised on learning weights independently on each CNN branch of a Triplet Network, producing a *multi-branch* architecture [27]. Hence, in what follows, we will use the two top-performing solutions in [27] as a baseline to evaluate the object recognition performance of solutions which are purely based on Machine Learning.

## 2.2. Hybrid Methods for Object Recognition

Broadly speaking, *hybrid reasoning methods* combine knowledge-based reasoning with ML. A detailed survey of hybrid methods developed to interpret images can be found in [10, 11]. Many of these hybrid methods are specifically tailored on Deep NNs, which define the predominant approach to tackling object recognition problems. In this setup, background knowledge and knowledge-based reasoning can be integrated at four different levels of the NN [10]: (i) in **pre-processing**, to augment the training examples, (ii) within the **intermediate layers**, (iii) as part of the **architectural topology** or **optimisation function**, and (iv) in the **post-processing** stages, to validate the NN predictions.

Methods in the first group rely on external knowledge to compensate for the lack of training examples. In [23], auxiliary images depicting newly-encountered objects were first retrieved

from the Web and then manually validated. As a result, significant supervision costs were introduced to compensate for the noisiness of data mined automatically. Other approaches have encoded the background knowledge directly in the inner layer representations of a Deep NN. In the RoboCSE framework, a set of knowledge properties of objects, (i.e., their typical location, fabrication material and affordances) were represented through multi-relational embeddings [28]. This method was proven effective to infer the object's material, location and affordances from its class, but performed poorly on object categorisation tasks (i.e., when asked to infer the object's class from its properties).

More transparent and explainable than multi-relational embeddings, methods in the third group are either inspired by the topology of external knowledge graphs [29] or introduce reasoning components which are trainable end-to-end [30, 31, 32]. Graph Search Neural Networks (GSNN) [29] resemble an input knowledge graph. In Logic Tensor Networks (LTN) [30], entities are represented as distinct points in a vector space, based on a set of soft-logic assertions linking these entities. In this framework, symbolic rules (which may adhere to probabilistic logic [31]) are added as constraints to the NN optimisation function. Similarly, in [32], differentiable knowledge statements (expressed in fuzzy logic) contribute to the training loss.

Finally, the fourth family of hybrid methods uses knowledge-based reasoning to validate the object predictions generated through ML. In [33] the ML predictions are first associated with the Qualitative Spatial Relationships in the image, and also matched against the top-most related DBpedia concepts, if an unknown object is observed. As in the case of [32], methods in this group can modularly interface with different NN architectures. Moreover, they make it possible to reason on objects unseen at training time, by querying external knowledge sources. For these reasons, in the approach proposed in this paper, knowledge-based reasoning is applied after generating the ML predictions. Because our focus is on reasoning about size, we cannot directly compare our approach against methods in [33], which focus on spatial and taxonomic reasoning (i.e., reasoning about the semantic relatedness of different object categories).

### 2.3. Representing the Size of Objects

Cognitive studies [13, 14] have suggested that our brain maintains a *canonical* representation of the physical size of objects, which is functional to their categorisation. Specifically, this prototypical size appears to be proportional to the logarithm of the known object size [13]. Successive experiments [14] have indicated that size-related features are extracted since the earliest visual processing stages, before object recognition. As such, it appears that a mid-level representation is produced to link the lower-level perceptual stimuli of our vision system to higher-level semantic concepts - e.g., our naming of objects. Remarkably, this representation of size is robust to variations in the shape and appearance of objects within a class, i.e., the *contour variance* of [14]. For example, a short novel and a dictionary are both books, although dictionaries are usually thicker. Moreover, the book may be open or closed, thus exhibiting a different size.

Works in Artificial Intelligence have sought inspiration from these cognitive theories. Based on the findings of [13], authors in [19] developed a reasoner where object sizes are represented as log-normal distributions over quantitative size measurements. Additionally, the produced distributions were used to populate nodes in a *size graph*, where objects which co-occur fre-

quently across different images are linked together. Similarly, in [20], the size of an object class was modelled quantitatively as a statistical distribution over a set of textual references to size. The size representations in [19, 20] were implemented to tackle textual reasoning problems, i.e., autonomously answer questions such as "is A larger than B?". In the context of autonomous visual reasoning, however, the effects of integrating size awareness remain unexplored. A partial exception is the work in [34], which proposes a methodology to build a Knowledge Base of object *affordances* (uses or actions typically associated with an object). As the main focus of [34] is affordance reasoning rather than pure size reasoning, however, the object size was only partially represented, by resorting to the object length derived from a combination of Freebase [35], Amazon and Ebay. All the reviewed representations [19, 20, 34] were extracted from repeated measurements retrieved from the Web. This approach, one the one hand, increases the chances to capture the contour variance within a class, because it takes multiple input measurements into account. Moreover, it reduces the cost of hardcoding a size Knowledge Base, especially given the lack of comprehensive resources encoding size, as pointed out in [19] and also highlighted by our prior KB coverage study [4]. On the other hand, however, this method is sensitive to noise. Another limitation of the reviewed methods is that size is represented in one-dimensional terms: e.g., either over volume or over length units. While relying on a single size feature is sufficient to identify broader groups of smaller and larger objects, it does not suit the task of classifying finer-grained object categories. For instance, recycling bins and coat stands may occupy a comparable volume, but bins are usually thicker than coat stands. Hence, while abstracting from lower-level size features, we also want to preserve enough information to categorise objects, i.e., produce the mid-level representation envisioned in [14]. To this aim, we propose to represent size qualitatively across three dimensions, i.e., based on the object surface area, depth and aspect ratio. To control for noise, we start by collecting these coarse-grained annotations manually, as further discussed in the next Section.

# 3. Methodology

## 3.1. Representing Qualitative Sizes in a Knowledge Base

We identified 60 object categories which are commonly found in KMi, the setting in which we aim to deploy our robotic Health and Safety monitor, HanS. These include not only objects which are common to most office spaces (e.g., chairs, desktop computers, keyboards), but also Health and Safety equipment (e.g., fire extinguishers, emergency exit signs, fire assembly point signs), and other miscellaneous items (e.g., a foosball table, colorful hats from previous gigs of the KMi rock band). The objective was then to label each object category with respect to a set of coarse-grained features contributing to size, namely their (i) **front surface area** (i.e., the product of their width by their height), (ii) **depth** dimension, and (iii) **Aspect Ratio (AR)**, i.e., the ratio of their width to their height. With respect to the first dimension, we can characterise objects as *extra-small*, *small*, *medium*, *large* or *extra-large* respectively. Secondly, objects can be categorised as *flat*, *thin*, *thick*, or *bulky*, based on their depth. Thirdly, we can further discriminate objects based on whether they are *taller than wide* (*ttw*), *wider than tall* (*wtt*), or *equivalent* (*eq*), i.e., of AR close to 1.

If the first two qualitative dimensions were plotted on a cartesian plane, a series of quadrants

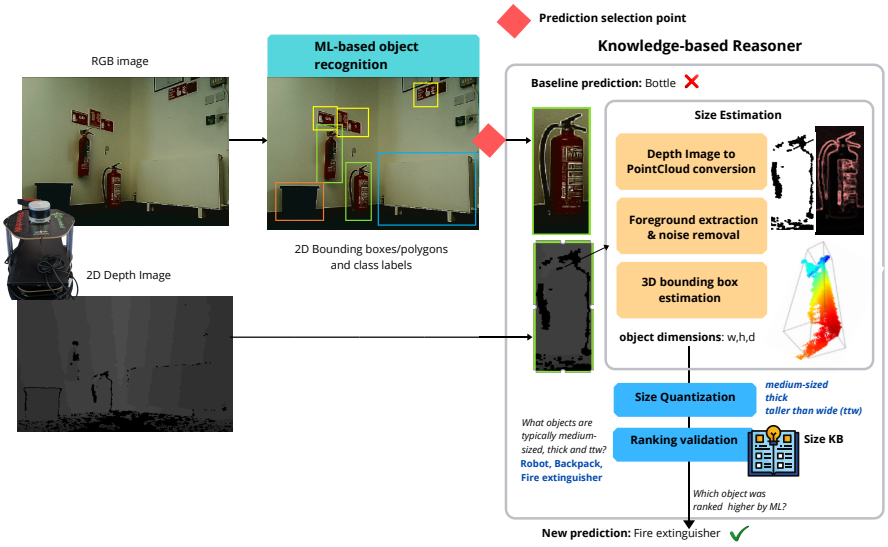


**Figure 1:** The proposed architecture for hybrid object recognition. The knowledge-based reasoning module, which is aware of the typical size features of objects, validates the ML-based predictions.

would emerge, as illustrated in this Figure: https://robots.kmi.open.ac.uk/img/size_repr.pdf. Then, the AR can help to further separate the clusters of objects belonging to the same quadrant. For instance, doors and desks both belong to the extra-large and bulky group but doors, contrarily to desks, are usually taller than wide. Having defined the cartesian plane in the supporting materials, we can manually allocate the KMi objects to each quadrant and further sort the objects lying in the same quadrant. Sorting the objects manually ensures more reliable results than if the same information was retrieved automatically.

Moreover, in the proposed representation, membership of each bin is mutually non-exclusive. Thus, with this representation, even classes which are extremely variable with respect to size, such as carton boxes and power cords, can be modelled. Indeed, boxes come in all sizes and power cords come in different lengths. Moreover, a box might lay completely flat, or appear bulkier, once assembled. Similarly, power cords, which are typically thinner than other pieces of IT equipment, might appear rolled up or tangled.

## 3.2. Hybrid Reasoning Architecture

We propose a modular approach to combining knowledge-based reasoning with Machine Learning for object recognition, where knowledge of the qualitative size of objects (Section 3.1) is integrated in post-processing, after generating the ML predictions. The architectural components are organised as follows (Figure 1).

**ML-based object recognition.** We rely on the state-of-the-art, ML-based object recognition methods of [27], to classify a set of pre-segmented object regions. Specifically, we classify objects by similarity to a reference image set, through a multi-branch Network. In this deep metric learning setting, predictions are ranked by increasing Euclidean (or L2) distance between each target embedding and the reference embedding set. Nevertheless, this configuration can

be easily replaced by any other algorithm that provides, for each detected object, (i) a set of class predictions with an associated measure of confidence (whether similarity-based or probability-based) and (ii) the segmented image region enclosing the object.

**Prediction selection.** This checkpoint is conceived for assessing whether a ML-based prediction needs to be corrected or not. At the time of writing, we achieved good results simply by retaining those predictions which the ML algorithm is most confident about, and by running the remaining predictions through the size-based reasoner. Specifically, we avoid the knowledge-based reasoning steps if the top-1 class in the ML ranking: (i) has a ranking score smaller than $\epsilon$ (i.e., the test image embedding lies near one of the reference embeddings, in terms of L2 distance); and also (ii) appears at least $i$ times in the top-K ranking.

**Size estimation.** At this stage, the input depth image corresponding to each RGB scene is first converted to a 3D PointCloud representation. Then, statistical outliers are filtered out to extract the dense 3D region which best approximates the volume of the observed object. Specifically, all points which lie farther away than two standard deviations ($2\sigma$) from their $n$ nearest neighbours are discarded. Because this outlier removal step is computationally expensive, especially for large 3D regions, we only retain 1 every $\chi$ points from the input PointCloud. We use the Convex Hull algorithm to approximate the 3D box bounding of the object and thus estimate its x,y,z dimensions. Since the orientation of the object is not known a priori, we cannot unequivocally map any of the estimated dimensions to the object's real width and height. However, we can assume the object's depth to be the minimum of the returned x,y,z dimensions, due to the way depth measurements are collected through the sensor. Indeed, since we do not apply any 3D reconstruction mechanisms, we can expect that the measured depth underestimates the real depth occupied by the object.

**Size quantization.** The three dimensions obtained at the previous step are here expressed in qualitative terms, to make them comparable with the representation of Section 3.1. First, the two dimensions which were not marked as depth are multiplied together, yielding a proxy of the object's surface area. The object is then categorised as extra-small, small, medium, large or extra-large, based on a threshold set $T$. Second, the estimated depth dimension is labelled as flat/non-flat (based on a threshold $\lambda_0$ ), and as flat, thin, thick, or bulky (based on a second set of thresholds $\Lambda$, where $\lambda_0 \in \Lambda$). Third, hypotheses are made about whether the object is taller than wide (ttw), wider than tall (wtt), or equivalent (eq), based on a cutoff $\omega_0 \in \Omega$. It would be unfeasible to predict the object's Aspect Ratio from the estimated 3D dimensions, without knowing its current orientation. Thus, we estimate the object's AR based on the width (w) and height (h) of the 2D bounding box:

$$AR = \begin{cases} ttw & \text{if } h \geq w \wedge \frac{h}{w} \geq \omega_0 \\ wwt & \text{if } h < w \wedge \frac{w}{h} \geq \omega_0 \\ eq & \text{otherwise} \end{cases} \tag{1}$$

**Ranking validation.** The qualitative features returned by the size quantization module are then matched against the background KB of Section 3.1, to identify the set of categories which are plausible candidates for the observed object, based on their size. In Section 4, we test the effects of combining different size features (i.e., the area surface, thinness, and AR) to generate this candidate set. Ultimately, only those object classes in the original ML ranking which were

validated as plausible are retained.

# 4. Experiments

## 4.1. Data Preparation

**KMi RGB Reference Set**. For training purposes, RGB images depicting the 60 target object classes were collected through a Turtlebot mounting an Orbbec Astra Pro RGB-Depth (RGB-D) monocular camera. Each object was captured against a neutral background and opportunely cropped to control for the presence of clutter and occluded parts. We collected 10 images per class, with an 80%-20% training-validation split. Specifically, 5 images per class were used as anchor examples and paired up with their most similar image among the remaining 5 examples in that class (i.e., the multi-anchor switch strategy in [27]). We also matched each anchor with the most similar image belonging to a different class. In this way, instead of picking negative examples randomly, i.e., the protocol followed in [27], we focused on triplets which are relatively harder to disambiguate.

**KMi RGB-D Test Set**. For testing purposes, additional RGB and depth measurements of the KMi office environment were collected, during one of HanS' monitoring routines. Class cardinalities in this set reflect the natural occurrence of objects in the observed domain - e.g., HanS is more likely to spot fire extinguishers than printers, on its scouting route. These data were recorded at a 640x480 resolution, i.e., the maximum resolution allowed by the depth sensor. 622 clear RGB frames, where the robot camera was still, were manually selected from the original recording. Because the recording of RGB and depth messages is asynchronous, each RGB frame had to be matched to its nearest depth image, within a time window of +/- $\mu$. Choosing a higher value for $\mu$ increases the number of scenes for which a depth match is available, but also increases the chances that a certain scene is matched with the wrong depth measurements (e.g., because the robot has moved in the meantime). In our trials, we found that setting $\mu$ to 0.2 seconds offered a good compromise. With this setup, a depth match was found for 509 ( 82%) of the 622 original frames. After a visual inspection, only 2 (0.4%) of the 509 depth matches were identified as inaccurate and discarded. This RGB-D set was further pruned to exclude identical images, where neither the robot's viewpoint nor the object arrangement had changed. We annotated the remaining 213 images with respect to 60 reference object categories. Concurrently, we also labelled the rectangular or polygonal regions bounding the objects of interest. The annotated regions were used to crop both the RGB frames and their time-synchronised depth images. Upon analysing the generated depth crop, we identified 19 object regions which did not enclose any depth measurement. This can happen, for instance, when the object falls outside the range of the depth sensor (i.e., 60 cm – 8 m). To fairly compare the performance of the size-based reasoner (which relies on depth data) against the ML baseline, we discarded the empty depth crops from our test set, leaving us with a total 1414 object regions.

## 4.2. Ablation Study

In what follows, we list the different methods under evaluation and illustrate the changes introduced before each performance assessment.

**Baseline Nearest Neighbour (NN).** In this pipeline, feature vectors are extracted from a ResNet50 module pre-trained on ImageNet [22], without re-training on the KMi RGB reference set. Namely, a 2048-dimensional embedding is extracted for each object region in the KMi RGB-D test set and matched to its nearest embedding in the KMi RGB reference set, in terms of L2 distance. This baseline provides us with a lower bound for the ML performance, before fine-tuning on the domain of interest.
**N-net** is the multi-branch Network which ensured the top performance on novel object classes, i.e., classes unseen at training time, in [27]. Training hyperparameters are updated so that the Triplet Loss is minimised, i.e., to minimise the L2 distance between matching pairs, while also maximising the L2 distance between dissimilar pairs. At inference time, object regions in the KMi RGB-D test set are classified based on their nearest object in the KMi RGB reference set, as in the case of the baseline NN pipeline.
**K-net** is the multi-branch Network which led to the top performance on known object classes, i.e., classes seen at training time, in [27]. K-net is a variation of N-net where a second loss component is added to the Triplet Loss. This auxiliary component of the loss derives from applying a SoftMax function over M known classes to the last fully connected layer.
**Hybrid (area).** This configuration follows the hybrid architecture introduced in Section 3.2. However, only the object's surface area is used as size feature to validate the ML predictions.
**Hybrid (area + flat).** With this ablation, we evaluate the effects of introducing a second feature to represent the size of objects. Specifically, here we consider not only the qualitative surface area of each object, but also whether they are flat or non-flat, based on their estimated depth. Then, the ML-based predictions are validated based on the set of object classes which both lie within the same area range and are also equally flat (or non-flat).
**Hybrid (area + thin)** is equivalent to the previous configuration, except the depth of objects is represented on a four-class scale: i.e., as flat, thin, thick, or bulky. The purpose of this ablation is testing the utility of introducing more granular depth bins.
**Hybrid (* + AR).** These ablations also integrate the qualitative Aspect Ratio (i.e., taller than wide, wider than tall, or equivalent) as a third knowledge property of size.

### 4.3. Implementation Details

**ML setup.** All the tested ML models were implemented in PyTorch [36]. Images in the KMi RGB reference set were resized to 224 × 224 frames and normalised to the same distribution as the ImageNet-1000 dataset, which was used for pre-training the ResNet50 CNN backbone. The tested ablations were fine-tuned through an Adabound optimizer [37], over minibatches of 16 image triplets, with a learning rate set to start at 0.0001 and to trigger switching to Stochastic Gradient Descend optimization at 0.01. Parameters were updated for up to 1500 epochs, with an early stopping whenever the validation loss had not decreased for more than 100 epochs.
**Knowledge-based reasoning parameters.** We relied on the Python Open3D library [38] to process the PointCloud data and estimate the bounding 3D rectangles. During our trials we achieved the best results with threshold values set as follows. In the prediction selection step, $\epsilon$ was set to a distance of 0.04 and $i$ to 3 predictions, within a top-5 ranking. The three cutoff sets in the size quantization module were defined so that $T$ ={0.007,0.05,0.35,0.79} (with thresholds expressed in squared meters), $\Lambda$ ={0.1,0.2,0.4} (in meters) and $\Omega$={1.4 times}.

**Table 1**
ML baseline results on the KMi RGB-D test set.

| | | Top-1 unweighted | | | Top-1 weighted | | | Top-5 results unweighted | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | Top-1 Acc. | P | R | F1 | P | R | F1 | P@5 | nDCG@5 | Hit ratio |
| Baseline NN | .33 | .36 | .33 | .25 | .63 | .33 | .36 | .23 | .25 | .54 |
| N-net | .45 | .34 | **.40** | .31 | .62 | .45 | .47 | .33 | .36 | .63 |
| K-net | **.48** | **.39** | **.40** | **.34** | **.68** | **.48** | **.50** | **.38** | **.41** | **.65** |

**Table 2**
Hybrid reasoning results, when correcting only the wrong predictions.

| | | Top-1 unweighted | | | Top-1 weighted | | | Top-5 results unweighted | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | Top-1 Acc. | P | R | F1 | P | R | F1 | P@5 | nDCG@5 | Hit ratio |
| Hybrid (area) | .60 | .52 | .50 | .47 | .71 | .60 | .61 | .43 | .47 | .72 |
| Hybrid (area+flat) | .60 | .55 | .50 | .48 | .72 | .60 | .62 | .44 | .47 | .72 |
| Hybrid (area+thin) | .61 | .55 | .50 | .48 | .71 | .61 | .63 | .44 | .47 | .71 |
| Hybrid (area+flat+AR) | **.62** | .59 | .50 | .49 | **.76** | **.62** | **.65** | **.45** | **.49** | **.75** |
| Hybrid (area+thin+AR) | **.62** | **.62** | **.51** | **.52** | .74 | **.62** | **.65** | **.45** | .48 | .74 |

As an additional output of this work, we have also publicly released the RGB-D image set, annotated knowledge properties, and code used in these experiments: https://github.com/kmi-robots/object_reasoner.

## 5. Results and Discussion

We measured performance on the KMi RGB-D test (i) in terms of the cross-class Accuracy, Precision (P), Recall (R) and F1 score of predictions in the top-1 result of the ranking; as well as (ii) based on the top-5 predictions in the ranking, in terms of mean Precision (P@5), mean normalised Discounted Cumulatve Gain (nDCG@5) and hit ratio (i.e., the ratio of the number of times the correct prediction appeared in the top-5 ranking to the total number of predictions). Specifically, the P, R and F1 metrics were aggregated across classes before and after weighing the averages by class support (i.e., based on the number of ground truth instances in each class). Measures@5 are unweighted. When comparing the different methods under evaluation, we prioritise improvements on the weighted F1 score, which accounts for the naturally imbalanced occurrence of classes in the test set. Moreover, at comparable top-1 results, we favour methods which provide higher quality top-5 rankings. Indeed, if the correct class was not ranked first but still appeared in the top-5 ranking, it would be easier for an additional reasoner (or human oracle) to correct the prediction.

First, we evaluated all methods which are purely based on Machine Learning, i.e., before any background knowledge about the typical size of objects is integrated. The results of this first assessment are reported in Table 1. K-net is the ML baseline which led to the top performance, across all evaluation metrics. Therefore, we considered the K-net predictions as a baseline, for testing all the hybrid configurations.

**Table 3**
Hybrid reasoning results, when correcting only an automatically selected subset of predictions.

| | | Top-1 unweighted | | | Top-1 weighted | | | Top-5 results unweighted | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | Top-1 Acc. | P | R | F1 | P | R | F1 | P@5 | nDCG@5 | Hit ratio |
| Hybrid (area) | **.55** | .42 | **.41** | **.38** | .67 | **.55** | **.57** | **.43** | **.46** | **.69** |
| Hybrid (area+flat) | .54 | **.43** | .39 | .37 | .67 | .54 | .56 | **.43** | **.46** | .68 |
| Hybrid (area+thin) | .52 | .39 | .37 | .35 | .62 | .52 | .54 | .42 | .44 | .64 |
| Hybrid (area+flat+AR) | .53 | **.43** | .37 | .36 | **.69** | .53 | .56 | **.43** | .45 | .68 |
| Hybrid (area+thin+AR) | .51 | **.43** | .36 | .36 | .64 | .51 | .54 | .41 | .43 | .63 |

Because the knowledge-based reasoner relies on different sub-modules, and each one of these modules is likely to propagate its own errors, we initially tested performance assuming that the ground truth predictions are known and that we can accurately discern which ML predictions need to be corrected. Although unrealistic, this best-case scenario provides us with an upper bound for the reasoner's performance and aids the analysis of errors. As shown in Table 2, simply integrating knowledge about the qualitative surface area of objects already ensured a significant performance improvement, with a **13%** increase of the unweighted F1 score and a **11%** increase of the weighted F1 score. Overall, the best performance, both in terms of top-1 predictions as well as in terms of top-5 rankings, was achieved through the two hybrid configurations which included all the qualitative size features (i.e., surface area, thinness and AR). In particular, the unweighted F1 score increased up to **18%** and the weighted F1 score up to **15%**. Hence, the margin for improvement when complementing ML with size-based reasoning is significant. These results confirm the hypothesis laid in [4]: the capability to compare objects by size and the access to background knowledge representing size play a crucial role in object categorisation. Notably, there is no significant difference between the results obtained when representing depth in binomial terms (i.e., as either flat or non-flat), as opposed to when more fine-grained categories are used (i.e., flat, thin, thick or bulky). Thus, we can hypothesise that the costs (and potential inaccuracies) associated with formalising additional priors for the objects' depth are not justified by a sufficient performance gain.

To capitalise on the latent performance gains highlighted in Table 2, the ML and knowledge based outcomes need to be opportunely leveraged. To this aim, we introduced a meta-reasoning checkpoint (i.e., the prediction selection module of Section 3.2) and automatically selected a subset of ML predictions to feed to the knowledge-based reasoner. The results of this last evaluation setup are summarised in Table 3. The ML baseline was outperformed by up to **4%** in terms of unweighted F1 and by up to **7%** in terms of weighted F1. Moreover, introducing knowledge about the object's surface positively impacted the quality of the top-5 ranking: the mean P@5 and nDCG@5 both increased by **5%**, and the hit ratio by **4%**. The qualitative surface area is the feature which led to the most consistent results across the different evaluation metrics. In other words, integrating additional knowledge beyond that first feature only led to comparable results, or even degraded the performance (i.e., in the case where a four-class scale instead of a binary one is used for the object's depth). Hence, in the experimental scenario of this paper, a size representation as minimalistic as indicating whether the object exposes an extra-small, small, medium, large, or extra-large front surface area is sufficient to ensure a

significant boost in performance.

## 6. Conclusion and Future Work

In this paper, we demonstrated that ML-based object recognition can be significantly augmented by a reasoning module which can account for the typical size of objects, as hypothesised in our prior work [4]. These results are particularly promising, because they were achieved on image regions collected by a robot in its natural environment, i.e., in a more challenging setup than benchmark image collections. In the proposed approach, we relied on a novel representation of the size of objects. Differently from prior knowledge representations, here we modelled size across three dimensions (the object's front surface area, depth and aspect ratio), to further separate the object clusters. Moreover, we allowed for annotating each object class with multiple size attributes, to adequately capture the size variability within each class.

The experiments presented in this paper also highlighted a series of directions of improvement, informing our future work. First, when estimating the object size from depth data we had to deal with hardware constraints: (i) objects falling outside the range of the depth sensors were excluded; (ii) highly reflective, absorptive or transparent materials (e.g., shiny metals, glass) altered the depth measurements. As such, access to a more advanced depth sensor would further improve the performance. Second, if the object was only partially visible in the original image, the estimated measurements (albeit accurate) would fail to represent the real object's size. Thus, incorporating the capability of moving towards the target object to refine the prediction through repeated measurements (i.e., Active Vision) is likely to benefit performance.

Naturally, the relevance of background knowledge and knowledge-based reasoning for enabling Visual Intelligence spans way beyond the capability to reason about the typical size of objects. In [4], we have identified several other reasoners (e.g., spatial, compositional, motion-aware) which may enhance the robustness of state-of-the-art Machine Learning methods. Thus, in our future work, we will evaluate the performance impacts of integrating: (i) additional knowledge-based components, (ii) multiple sources of background knowledge, as well as (iii) effective meta-reasoning strategies, to reconcile the outcomes of different reasoners.

## References


[1] M. Bajones, D. Fischinger, A. Weiss, D. Wolf, M. Vincze, de la Puente, et al., Hobbit: Providing Fall Detection and Prevention for the Elderly in the Real World, Journal of Robotics (2018).

[2] F. Dong, S. Fang, Y. Xu, Design and Implementation of Security Robot for Public Safety, in: 2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS), 2018, pp. 446–449.

[3] J. Waldhart, A. Clodic, R. Alami, Reasoning on Shared Visual Perspective to Improve Route Directions, in: 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2019, pp. 1–8.

[4] A. Chiatti, E. Motta, E. Daga, Towards a Framework for Visual Intelligence in Service

Robotics: Epistemic Requirements and Gap Analysis, in: Proceedings of KR 2020- Special session on KR & Robotics, IJCAI, 2020, pp. 905–916.
[5] 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.
[6] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015).
[7] G. Marcus, Deep learning: A critical appraisal, arXiv preprint arXiv:1801.00631 (2018).
[8] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, S. Wermter, Continual lifelong learning with neural networks: A review, Neural Networks 113 (2019) 54–71.
[9] J. Pearl, Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution, in: Proceedings of WSDM 2018, ACM, 2018, p. 3.
[10] S. Aditya, Y. Yang, C. Baral, Integrating Knowledge and Reasoning in Image Understanding, in: Proceedings of IJCAI 2019, 2019, pp. 6252–6259.
[11] F. Gouidis, A. Vassiliades, T. Patkos, A. Argyros, N. Bassiliades, D. Plexousakis, A Review on Intelligent Object Perception Methods Combining Knowledge-based Reasoning and Machine Learning, arXiv:1912.11861 [cs] (2020).
[12] A. A. Daruna, V. Chu, W. Liu, M. Hahn, P. Khante, S. Chernova, A. Thomaz, Sirok: Situated robot knowledge-understanding the balance between situated knowledge and variability, in: 2018 AAAI Spring Symposium Series, 2018.
[13] T. Konkle, A. Oliva, Canonical visual size for real-world objects, Journal of Experimental Psychology: Human Perception and Performance 37 (2011).
[14] 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.
[15] E. Davis, G. Marcus, Commonsense reasoning and commonsense knowledge in artificial intelligence, Communications of the ACM 58 (2015) 92–103.
[16] H. Levesque, Common Sense, the Turing Test, and the Quest for Real AI | The MIT Press, The MIT Press, 2017.
[17] P. J. Hayes, The Second Naive Physics Manifesto, Formal theories of the common sense world (1988). Publisher: Ablex Publishing Corporation.
[18] 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).
[19] H. Bagherinezhad, H. Hajishirzi, Y. Choi, A. Farhadi, Are elephants bigger than butterflies? reasoning about sizes of objects, in: Proceedings of AAAI, AAAI'16, AAAI Press, Phoenix, Arizona, 2016, pp. 3449–3456.
[20] Y. Elazar, A. Mahabal, D. Ramachandran, T. Bedrax-Weiss, D. Roth, How Large Are Lions? Inducing Distributions over Quantitative Attributes, in: Proceedings of the ACL, Association for Computational Linguistics, 2019, pp. 3973–3983.
[21] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, Communications of the ACM 60 (2017) 84–90.
[22] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proceedings of CVPR, 2016, pp. 770–778.
[23] M. Mancini, H. Karaoguz, E. Ricci, P. Jensfelt, B. Caputo, Knowledge is Never Enough: Towards Web Aided Deep Open World Recognition, in: IEEE ICRA, 2019, p. 9543.

[24] G. Koch, R. Zemel, R. Salakhutdinov, Siamese neural networks for one-shot image recognition, in: ICML deep learning workshop, volume 2, Lille, 2015.
[25] E. Hoffer, N. Ailon, Deep Metric Learning Using Triplet Network, Lecture Notes in Computer Science (2015) 84–92.
[26] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: A Unified Embedding for Face Recognition and Clustering, in: Proceedings of the IEEE CVPR, 2015, pp. 815–823.
[27] 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, IEEE, 2018, pp. 1–8.
[28] A. Daruna, W. Liu, Z. Kira, S. Chetnova, Robocse: Robot common sense embedding, in: Proceedings of ICRA, IEEE, 2019, pp. 9777–9783.
[29] K. Marino, R. Salakhutdinov, A. Gupta, The More You Know: Using Knowledge Graphs for Image Classification, in: Proceedings of IEEE CVPR, 2017, pp. 20–28.
[30] L. Serafini, A. d. Garcez, Logic Tensor Networks: Deep Learning and Logical Reasoning from Data and Knowledge, arXiv:1606.04422 [cs] (2016).
[31] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, L. De Raedt, Deepproblog: Neural probabilistic logic programming, in: Advances in Neural Information Processing Systems, 2018, pp. 3749–3759.
[32] E. van Krieken, E. Acar, F. van Harmelen, Analyzing Differentiable Fuzzy Implications, in: Proceedings of KR 2020, 2020, pp. 893–903.
[33] J. Young, L. Kunze, V. Basile, E. Cabrio, N. Hawes, B. Caputo, Semantic web-mining and deep vision for lifelong object discovery, in: Proceedings of ICRA, IEEE, 2017, pp. 2774–2779.
[34] Y. Zhu, A. Fathi, L. Fei-Fei, Reasoning about Object Affordances in a Knowledge Base Representation, in: Proceedings of ECCV, volume 8690, Springer International Publishing, 2014, pp. 408–424.
[35] 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.
[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, et al., PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8026–8037.
[37] L. Luo, Y. Xiong, Y. Liu, X. Sun, Adaptive Gradient Methods with Dynamic Bound of Learning Rate, in: Proceedings of ICLR, 2018.
[38] Q.-Y. Zhou, J. Park, V. Koltun, Open3d: A modern library for 3d data processing, arXiv preprint arXiv:1801.09847 (2018).