Towards Visually Intelligent Agents (VIA): a Hybrid Approach

Agnese Chiatti[1]
Middle stage PhD Student[orcid=0000−0003−3594−731X]

Knowledge Media Institute, The Open University,
Walton Hall, Milton Keynes, MK7 6AA, United Kingdom
agnese.chiatti@open.ac.uk

Abstract. Service robots can undertake tasks that are impractical or even dangerous for us - e.g., industrial welding, space exploration, and others. To carry out these tasks reliably, however, they need Visual Intelligence capabilities at least comparable to those of humans. Despite the technological advances enabled by Deep Learning (DL) methods, Machine Visual Intelligence is still vastly inferior to Human Visual Intelligence. Methods which augment DL with Semantic Web technologies, on the other hand, have shown promising results. In the lack of concrete guidelines on which knowledge properties and reasoning capabilities to leverage within this new class of hybrid methods, this PhD work provides a reference framework of epistemic requirements for the development of Visually Intelligent Agents (VIA). Moreover, the proposed framework is used to derive a novel hybrid reasoning architecture, to address real-world robotic scenarios which require Visual Intelligence.

Keywords: Hybrid AI · Visual Intelligence · Service Robotics.

1 Introduction and Motivation

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, especially in scenarios where it is unsafe or impractical for us to intervene - e.g., under extreme weather conditions or when social distance needs to be maintained. However, succeeding in the real world is a challenge because it requires robots to make sense of the high-volume and diverse data collected through their perceptual sensors [2]. From the entry point of vision, in particular, 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* [8]. The first prerequisite to building *Visually Intelligent Agents (VIA)* is the ability to robustly *recognise* the different *objects* occupying the robot's environment. Let us consider the case of HanS, the Health and Safety (H&S) robot inspector at the Knowledge Media Institute (KMi) [4]. 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 first needs to detect the cup and the heater. However, HanS also needs access to many other reasoning capabilities and knowledge components: it needs to know that paper cups are flammable, and that portable heaters can produce heat. It also needs spatial reasoning capabilities, to infer that the cup is touching the heater, and so forth.

Currently, the predominant approach to tackling visual reasoning 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) [22, 20]. Despite their popularity, these methods have received many critiques due to their brittleness and lack of transparency [24, 26, 28]. These limitations are particularly evident when compared against the excellence of the human vision system [19, 15]. Indeed, we can learn rich object representations very rapidly, even from minimal observations, and adapt these representations to reflect changes in the environment. To compensate for the limitations of ML-based methods, a more recent trend among AI researchers has been to combine ML with knowledge-based reasoning, thus adopting a *hybrid approach* [1, 13]. Concurrently, thanks to efforts in the Semantic Web and Knowledge Engineering communities, an increasing number of large-scale resources encoding linguistical, encyclopaedical and common-sense knowledge have been made available [30]. Thus, a promising research direction is capitalising on these knowledge resources to develop hybrid reasoning architectures. A question remains, however, on what type of knowledge resources and reasoning capabilities should be leveraged within hybrid methods [11].

Based on these premises, the first objective of this PhD research is identifying a set of *epistemic requirements*, i.e., a set of capabilities and knowledge properties, required for service robots to exhibit Visual Intelligence. Another objective is mapping these epistemic ingredients to the knowledge properties available within state-of-the-art Knowledge Bases (KB), to evaluate to which extent they can support VIA. Together, the produced requirement analysis and coverage study provide a framework for the development of VIA which is fit for use, as well as a research agenda to build improved knowledge representations for robotic applications. Moreover, the error analysis informs our hypotheses on which epistemic requirements to prioritise, in the real-world use-case of monitoring H&S in the office. Specifically, our intermediate results [8, 9] indicate that knowledge of the typical size of objects and of their typical spatial locations are key factors contributing to Visual Intelligence. Thus, in this work, a hybrid architecture is proposed, which leverages both types of reasoners.

This paper is structured as follows. Section 2 reviews the state of the art in autonomous reasoning for Visual Intelligence. The research questions informing this work are presented in Section 3. Section 4 describes the methodological rationale followed to tackle each of these questions. Additionally, the proposed experimental design plan is discussed in Section 5. The proposal concludes with overviewing the current research progress as well as the next relevant activities.

2 Summary of Literature Review

Machine Learning methods (and the Deep Learning paradigm in particular) have expedited the improvement on several Computer Vision benchmarks [22, 18, 14]. Deep Neural Networks (NNs), however, come with their limitations. These models (i) are notoriously data-hungry, (ii) assume to operate in a closed world [23], and (iii) extract representational patterns through successive iterations over raw data [20]. 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 [24, 28]. Considering the limitations of state-of-the-art visual reasoning methods based on ML, hybrid approaches to visual reasoning, i.e., methods which combine ML with knowledge-based components, have been recently proposed [1, 13]. In DL setups, in particular, knowledge-based reasoning can be integrated at four different levels of the NN [1]: (i) in **pre-processing**, to augment the training examples [23], (ii) within the **intermediate layers** [10], (iii) as part of the **architectural topology** or **optimisation function** [25, 29, 16] , and (iv) in the **post-processing** stages, to validate the NN predictions [33]. Compared to the other classes of hybrid methods, a post-hoc approach offers the advantage of modularity, i.e., it is agnostic to the specific ML architecture used. Additionally, this approach increases the transparency of results, because it allows to decouple the ML predictions from the knowledge-based predictions and, thus, to evaluate how the different architectural components contribute to the overall performance. This characteristic is an important pre-condition to identifying the strengths, weaknesses and complementarities of each module, so that a more seamless integration is ensured and potentially conflicting outcomes between the different ML-based and knowledge-based predictors are handled effectively. For instance, on the one hand, applying off-the-shelf DL-based methods typically allows faster inference at test time than querying various knowledge sources [19]. On the other hand, the integration of large-scale knowledge bases allows a more transparent control of which knowledge properties and features contribute to the reasoning process. Thus, a hybrid system is expected to capitalise on the best of both worlds. Nonetheless, the literature lacks a systematic study of which ML-based and knowledge-based components are to be leveraged in hybrid systems. Specifically, this PhD work is focused on approaching this open problem from the angle of improving the Visual Intelligence of robots to support real-world application scenarios.

With the evolution of Semantic Web technologies, many large-scale knowledge resources have become available, which can be integrated within hybrid frameworks, such as the knowledge representations surveyed in [27, 30, 31, 21]. However, because several different types of background knowledge and reasoning capabilities are needed for robots to exhibit Visual Intelligence, choosing which knowledge resources and reasoning components to prioritise within hybrid architectures remains an open problem [11]. In [8] we have analyzed the types of classification errors emerging during robot monitoring activities, after applying state-of-the-art ML methods. Our error analysis indicated that two epistemic components, in particular, have the potential to significantly improve

the robot's capability to recognise objects: (i) the ability to compare objects by size, (ii) qualitative spatial reasoning capabilities. Indeed, the intermediate results of this PhD work [9] show that a novel hybrid system where knowledge of the typical size of objects is integrated in post-processing can significantly augment object recognition pipelines which are purely based on ML. With respect to the implementation of spatial reasoning capabilities, we propose a novel framework for qualitative spatial reasoning, which extends the work in [12, 5]. Differently from existing approaches, the proposed approach provides a mapping between formal representations of space in AI and the types of commonsense spatial representations used in everyday language [3]. As such, the proposed representational framework can be used to extract commonsense Qualitative Spatial Relations (QSR) from large-scale KBs which encode spatial knowledge [31, 21, 30]. Crucially, the proposed mapping can be fully implemented with state-of-the-art Geographic Information System (GIS) technologies.
Overall, the results obtained from evaluating the two proposed reasoners will inform the implementation of a meta-reasoning architecture, which can exploit the complementary strenghts of the ML-based and knowledge-based reasoners.

3 Problem Statement and Contributions

The main objective of this doctoral research is to study ways to improve the Visual Intelligence of service robots when making sense of complex, real-world environments. Based on evidence from the literature, the overarching hypothesis is that: *A hybrid approach (ML-based and knowledge-based) can improve a robot's performance on tasks that require Visual Intelligence (e.g., sensemaking), compared to approaches which rely solely on Machine Learning techniques.*

This hypothesis also raises a series of research questions. First, **RQ1:** *what are epistemic requirements, i.e., the set of required knowledge components and reasoning capabilities, of developing Visually Intelligent Agents?* Second, **RQ2:** *which epistemic requirements are the most important ones, in the considered use-case scenario?* Specifically, the intermediate results achieved while tackling RQ2 have indicated that two epistemic requirements, in particular, have the potential to significantly enhance HanS' Visual Intelligence: (i) the capability to reason on the physical size of objects, and (ii) the capability to reason about the spatial relations between objects. Therefore, the further inquiry will focus not only on *the extent to which the state-of-the-art Knowledge Bases support VIA* **(RQ3)**, but also on *the extent to which existing resources can be repurposed to support size and spatial reasoning* **(RQ4)**. Hence, another related question is about *the extent to which a concrete architecture which effectively leverages both types of reasoners can be developed* (**RQ5**).

4 Research Methodology

To address RQ1, requirements are gathered both through a top-down approach, i.e., based on seminal frameworks describing the human visual cognition, and

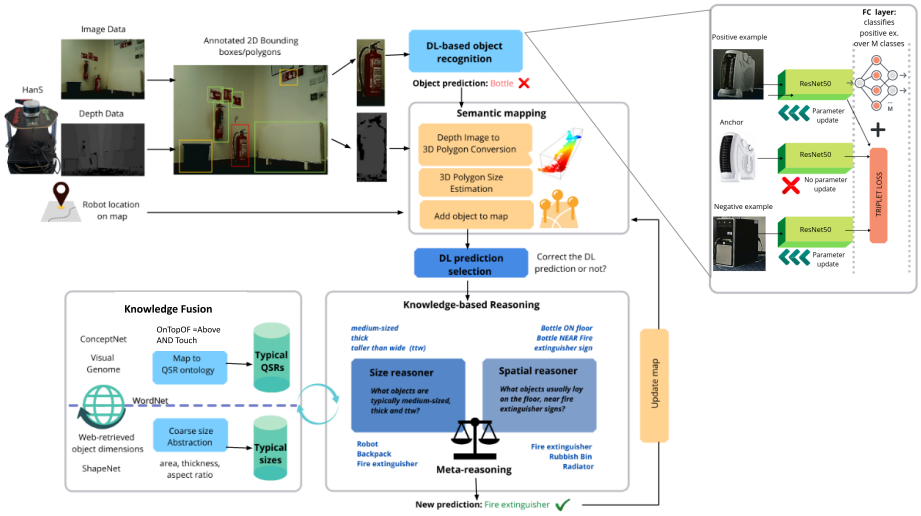

Fig. 1. The proposed hybrid architecture which leverages size and spatial reasoning. In this instance, hybrid reasoning is applied to the case of object recognition tasks.

from the bottom-up, i.e., based on the errors emerged from a real-world application scenario. The incentive of taking inspiration from the Human Visual Intelligence is motivated by the brittleness of current approaches to Machine Visual Intelligence. In addition to cognitively-inspired requirements, however, concrete requirements gather from error analysis are included as well. As such, the error analysis also provides a way to asses the relative impact of each requirement (RQ2). Moreover, the identified epistemic requirements can be used to assess the coverage of each required knowledge component that is provided with the state-of-the-art KBs identified in Section 2.3 (RQ3).

The results from RQ2 and RQ3 inform the selection of which reasoners and external KBs to include in a concrete hybrid architecture for VIA, also exemplified in Figure 1. The proposed architecture integrates auxiliary knowledge in post-processing, i.e., after generating the ML-based predictions. The object recognition pipeline exemplified in Figure 1 relies on the state-of-the-art multi-branch Network of [34]. In this setup, the NN is optimised to learn a feature space where similar objects lie closer than dissimilar objects. Training triplets consist of an anchor image, a positive (similar) example to the anchor, as well as a negative (dissimilar) example. At inference time, for each observed object, a ranking of object predictions is produced, based on similarity matching against the learned image embeddings. A few-shot metric learning approach was chosen as ML baseline to keep the required training examples to a minimum, while also ensuring that objects unseen at training time can still be classified at test time, by matching the learned representations against a reference image set. Nevertheless, the hybrid approach proposed in this work is general and any ML-based

methods which provides the bounding boxes and predicted categories for the observed objects can modularly interface with it. The Knowledge Base supporting this reasoning architecture will include: (i) a novel coarse-grained representation of size abstracted from lower-level size features, as further illustrated in [9], (ii) Qualitative Spatial Relations (QSR) gathered from a combination of general-purpose KBs, which are repurposed automatically through a dedicated knowledge fusion module (RQ4). Size and spatial knowledge is here represented qualitatively, to ensure the scalability of the proposed solution to broader application scenarios. A crucial component of the envisioned architecture is the meta-reasoning module, where the outcomes of different reasoners are opportunely leveraged, to converge towards a final set of object predictions. Therefore, a detailed ablation study will be carried out to identify the strengths and weaknesses of each component contributing to the overall performance. Indeed, the background knowledge available may be incomplete or unreliable. Similarly, the ML algorithm will be biased towards the patterns learned from the distribution of the training set. Thus, conflicting recommendations need to be leveraged, in an ensemble approach.

It is also worth noting that, although this PhD work is focused on implementing an architecture which combines size and spatial reasoning with ML, the proposed hybrid architecture is general, i.e., any other cognitive reasoner identified in [8] can be plugged in. Thus, in Figure 1, we use the broad term "Knowledge-based Reasoning" to refer to the process validating the knowledge properties extracted from the robot's observations against knowledge priors gathered from external resources. The size and spatial reasoner are only two instances of this general approach.

Another requirement to test the utility of the proposed architecture (RQ5) is defining a predetermined set of evaluation tasks that entail Visual Intelligence capabilities. These tasks are derived from the use-case scenario of H&S monitoring in the office. Namely, to anticipate the emergence of H&S threats through Vision, a robot will need to: (i) robustly recognise a set of known objects in a target environment (i.e., the task of *object recognition*), (ii) update its learning models and knowledge base, when exposed to new object classes (i.e., the task of *incremental object learning*), (iii) react based on the interpreted state of the environment - e.g., notify the designated fire wardens in case of a fire (i.e., *decision-making tasks*). An evaluation plan for each of these tasks is provided in the next Section.

5 Evaluation Plan

KB Evaluation Based on the epistemic requirements identified through RQ1, in [8], we have constructed a matrix where columns correspond to the identified knowledge requirements and rows indicate the state-of-the-art KBs reviewed in Section 2.3. The level of coverage of the required knowledge properties provided with each KB was then assessed on a qualitative scale.

Object Recognition The state-of-the-art ML methods presented in [34] were taken as baseline to conduct preliminary trials during the robot's patrolling rounds. A qualitative error analysis has been conducted on the basis of these preliminary data collection and trials, as further illustrated in [8]. Specifically, each classification error was recorded on a Boolean matrix, to mark the epistemic requirements which would have helped: (i) identifying the ground truth class, or (ii) ruling out the incorrect class. Then, in [9], the reference ML baselines were quantitatively evaluated on a larger dataset, to measure the performance effects of integrating knowledge of the typical size of objects. Performance was here evaluated based on: the P,R and F1 of the top-1 predictions; the standard ranking quality metrics P@5, Mean Normalised Discounted Cumulative Gain (Mean nDCG@5) and hit ratio. Specifically, the P, R and F1 were aggregated class-wise before and after weighing the averages by class support, i.e., the number of instances within each class, to account for the natural class imbalance in the dataset (e.g., fire extinguishers occur more often than printers, on the robot's scouting route). In these experiments, all object classes have been treated as known, i.e., introduced since training time. The same experimental setup will be replicated to test the introduction of the spatial reasoning module. Moreover, further tests will be conducted to evaluate the computational overhead introduced by the post-hoc reasoning steps, by tracking the processing times of each tested hybrid solution. Additional metrics to measure the inter-agreement (e.g., MCC and Cohen's Kappa) between the different ML-based and knowledge-based classifiers will be also considered, to inform the implementation of the meta-reasoning module.

Incremental Object Learning To test the scalability of the proposed hybrid framework to novel objects, i.e., unseen at training time, the first step has been to reproduce the experimental setup of the selected ML baselines [34]. In this setup, two ML-based methods are applied: (i) K-net, trained to overfit on a set of known objects, (ii) N-net, conceived to generalise to novel objects. Indeed, the dataset introduced in [34] in the context of the 2017 Amazon Robotic Challenge includes a combination of known objects, i.e., seen since training time, and novel objects, i.e., introduced only at test time. A preliminary ablation study on this datasets has allowed us to test performance in the presence of novel object classes. Because images in the Amazon dataset [34] only depict one object at a time, this dataset is not suitable for evaluating the performance of the spatial reasoner. Nonetheless, different splits of robot's dataset collected at the prior step will be tested, where only a subset of objects is treated as known.

Decision-making The objective of this phase is evaluating the robot's ability to reliably assess the state of risk of the environment it is monitoring. To this aim, a set of Health & Safety risk assessment scenarios will be defined: e.g., notifying fire wardens that a pile of paper was let on top of a portable heater or that the path of an emergency exit is not correctly signalled. In this phase, H&S experts at the Open University will be involved through a focus group discussion, to converge towards a small set of use-case scenarios which the experts consider as useful and worth implementing. Then, for each scenario, the robot's performance

will be evaluated based on: (i) the accuracy of the assessments (compared to the expert's indicated risk), (ii) the time elapsed before completing each assessment.

6 Summary of Intermediate Results

This Section summarises the current progress in tackling the research questions guiding this PhD research. Thus, in what follows, intermediate results are organised by research question.

RQ1: *what are the epistemic requirements of developing VIA?* In [8], we identified a set of top-down epistemic ingredients. Specifically, the requirement of learning as model building is transversal to all the other ingredients and entails: (i) defining concept representations and taxonomies which can be adequately expanded as new concepts are learned, as well as (ii) causal reasoning capabilities. The remaining top-down ingredients are: (iii) Intuitive Physics, (iv) compositionality, (v) Generic 2D views, (vi) Motion Vision, and (vii) fast perception. Thanks to a bottom-up analysis of object recognition errors emerging in a real-world robotic scenario, I have also completed the former set of requirements with (viii) the Machine Reading capability.

RQ2: *which epistemic requirements are the most important ones, in the considered use-case scenario?* The error analysis conducted in [8], also summarised in Figure 2, indicates that the majority of ML misclassifications could have been in principle avoided, with access to: (i) knowledge of the typical size of objects and the capability to compare objects by size, which falls under the Intuitive Physics component; (ii) knowledge of the typical Qualitative Spatial Relations (QSR) between objects, as well as spatial reasoning capabilities, which are part of the epistemic requirement of compositionality. Our most recent empirical findings also confirmed that size reasoning can significantly augment the object recognition performance of state-of-the-art ML solutions. On the KMi dataset (Table 1), the tested hybrid solution which integrates all the proposed size features (front surface area, thickness and Aspect Ratio) ensured to improve the unweighted and weighted F1 scores by **6%** and **5%**, compared to ML baselines. The quality of the top-5 results in the ranking also improved as a result of introducing these knowledge priors. Notably, in the case of the Amazon dataset, i.e., in the presence of known and novel objects, and of two ML algorithms of complementary efficacy, the introduced size reasoner provided a rationale to dynamically choose which ML algorithm to apply in each case. As highlighted in Table 2, the top-1 accuracy increased by **9.5%** in this scenario.

RQ3: *to what extent do the state-of-the-art Knowledge Bases support VIA?* In [8], we selected a set of KBs for review and assessed their coverage of the knowledge properties required for VIA. None of the reviewed KBs covers the identified knowledge requirements in full. The two most impactful knowledge attributes exposed by the bottom-up analysis (the object relative sizes and QSR) are covered only for a limited set of objects. Particularly striking is the lack of comprehensive knowledge representations which describe the typical motion trajectories of objects, e.g., as static or moving. Nonetheless, this coverage study

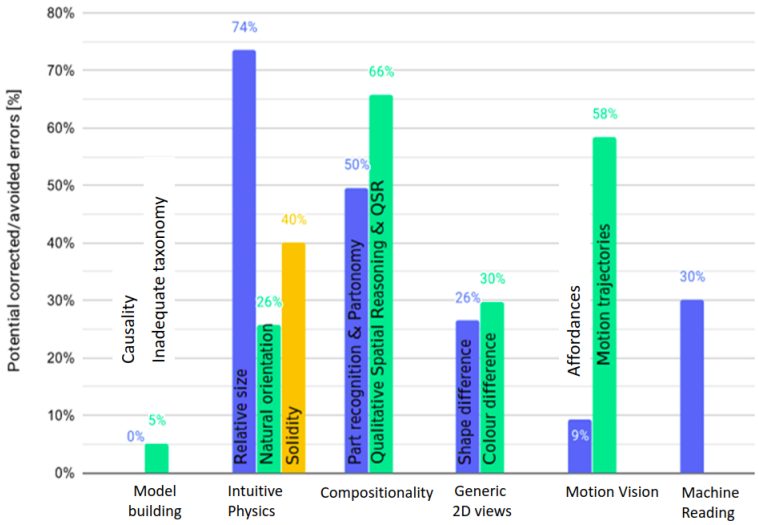

Fig. 2. From [8]: percentage of cases where a specific component of Visual Intelligence would help correcting or avoiding the classification error.

highlighted that most of the reviewed KBs are complementary to one another, with respect to the types of provided knowledge properties. Thus, a promising research direction is combining different external KBs to capitalise on synergistic effects.

RQ4: *to what extent the state-of-the-art Knowledge Bases be repurposed, to support size and spatial reasoning?* The positive performance results highlighted in Tables 1 and 2 were achieved thanks to automatically generating a catalogue of qualitative size descriptions from raw size measurements gathered from a combination of ShapeNet [6], Amazon and manual collection. Moreover, in [7], we have proposed a knowledge representation framework to map the commonsense and linguistic spatial predicates provided with state-of-the-art KBs to both: (i) the spatial operators available within state-of-the-art spatial databases, and (ii) formal AI statements expressed in First Order Logic (FOL). The next step will be applying the proposed framework to the extraction of spatial priors from general-purpose KBs such as Visual Genome [17], SpatialSense [32], and others.

7 Conclusions and Lessons Learned

Before we delegate complex tasks to robots, we need to ensure that they can reliably make sense of their environment. This PhD work proposes a framework of epistemic requirements for the development of Visually Intelligent Agents (VIA), i.e., robots which exhibit improved visual sensemaking capabilities. In particular, the main hypothesis underlying this work is that adopting a hybrid

	Top-1 unweigh.			Top-1 weigh.			Top-5 unweigh.		
Method	P	R	F1	P	R	F1	P@5	nDCG@5	HR
N-net [34]	34.0	**40.1**	31.0	61.5	45.2	47.2	33.1	36.0	63.0
K-net [34]	39.0	39.9	34.0	68.0	47.9	50.4	38.5	40.7	65.1
Hybrid (area)	39.6	39.5	35.5	65.5	50.3	51.6	41.0	43.1	68.0
Hybrid (area+flat/non-flat)	41.0	39.3	35.7	65.8	50.1	52.1	40.5	42.8	65.8
Hybrid (area+thickness)	44.5	38.9	38.6	65.0	**51.4**	53.9	**41.8**	**44.1**	**68.5**
Hybrid (area+flat/non-flat+AR)	42.9	38.8	36.6	68.9	49.1	52.9	39.9	42.0	66.3
Hybrid (area+thickness+AR)	**47.2**	39.1	**40.0**	**69.1**	**51.4**	**55.4**	41.6	43.9	68.4

Table 1: Evaluation results (in percentages), on the KMi test set.

	Top-1 accuracy			Top-5 unweighted		
Method	Known	Novel	Mixed	P@5	nDCG@5	HR
N-net [34]	56.8	**82.1**	64.6	61.9	62.7	72.6
K-net [34]	**99.7**	29.5	78.1	73.7	75.0	82.4
Hybrid (area)	94.7	71.7	**87.6**	**82.6**	**84.1**	**89.7**
Hybrid (area + flat/non-flat)	94.5	71.7	87.5	82.5	84.0	**89.7**
Hybrid (area + thickness)	81.7	39.3	68.7	64.6	65.8	70.1

Table 2: Evaluation results (in percentages), on the test set of [34].

approach, which combines Machine Learning with Semantic Web technologies, has the potential to significantly improve the performance of service robots on tasks that require Visual Intelligence. To test this hypothesis, a system is devised which integrates ML with two types of knowledge-based reasoners: (i) a reasoner which can take object sizes into account, and (ii) a qualitative spatial reasoner. The utility of this hybrid system is evaluated in the context of real-world robotic scenarios. At the time of this writing, the epistemic framework for VIA has already been defined and used to verify the level of support to the development of VIA which is provided with state-of-the-art Knowledge Bases [8]. Moreover, the intermediate results of this work show that a hybrid reasoner which integrates knowledge of the typical object sizes can significantly outperform object recognition methods based on ML. Nonetheless, the evaluation of the proposed framework is still incomplete, specifically with respect to testing: (i) the effects of integrating the spatial reasoning module presented in [7], (ii) the possibility to leverage both reasoners to reconcile potentially conflicting outcomes, (iii) the scalability to novel objects, as well as (iv) the level of support to concrete decision-making tasks which require Visual Intelligence.

Acknowledgements. I would like to thank my supervisors, Prof. Enrico Motta and Dr. Enrico Daga, for their continuous support and guidance throughout this PhD project. It is also thanks to them if I have found out about the ESWC PhD symposium.

References

1. Aditya, S., Yang, Y., Baral, C.: Integrating Knowledge and Reasoning in Image Understanding. In: Proceedings of IJCAI 2019. pp. 6252–6259 (2019)
2. Alatise, M.B., Hancke, G.P.: A review on challenges of autonomous mobile robot and sensor fusion methods. IEEE Access **8**, 39830–39846 (2020)
3. Barbara Landau, Jackendoff, R.: "What" and" where" in spatial language and spatial cognition. Behavioral and Brain Sciences **16**, 217–265 (1993)
4. Bastianelli, E., Bardaro, G., Tiddi, I., Motta, E.: Meet hans, the heath & safety autonomous inspector. In: Proceedings of the International Semantic Web Conference (ISWC), Poster&Demo Track (2018)
5. Borrmann, A., Rank, E.: Query Support for BIMs using Semantic and Spatial Conditions. In: Handbook of Research on Building Information Modeling and Construction Informatics: Concepts and Technologies (2010)
6. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
7. Chiatti, A., Bardaro, G., Motta, E., Daga, E.: Commonsense spatial reasoning for visually intelligent agents. arXiv preprint arXiv:2104.00387 (2021)
8. Chiatti, A., Motta, E., Daga, E.: Towards a Framework for Visual Intelligence in Service Robotics: Epistemic Requirements and Gap Analysis. In: Proceedings of KR 2020- Special session on KR & Robotics. pp. 905–916. Publisher: IJCAI (2020)
9. Chiatti, A., Motta, E., Daga, E., Bardaro, G.: 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)
10. Daruna, A., Liu, W., Kira, Z., Chetnova, S.: Robocse: Robot common sense embedding. In: Proceedings of ICRA. pp. 9777–9783. IEEE (2019)
11. Daruna, A.A., Chu, V., Liu, W., Hahn, M., Khante, P., Chernova, S., Thomaz, A.: Sirok: Situated robot knowledge-understanding the balance between situated knowledge and variability. In: 2018 AAAI Spring Symposium Series (2018)
12. Deeken, H., Wiemann, T., Hertzberg, J.: Grounding semantic maps in spatial databases. Robotics and Autonomous Systems **105**, 146–165 (Jul 2018)
13. Gouidis, F., Vassiliades, A., Patkos, T., Argyros, A., Bassiliades, N., Plexousakis, D.: A Review on Intelligent Object Perception Methods Combining Knowledge-based Reasoning and Machine Learning. arXiv:1912.11861 [cs] (Mar 2020)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of CVPR. pp. 770–778 (2016)
15. Hoffman, D.D.: Visual intelligence: How we create what we see. WW Norton & Company (2000)
16. van Krieken, E., Acar, E., van Harmelen, F.: Analyzing Differentiable Fuzzy Implications. In: Proceedings of KR 2020. pp. 893–903 (2020)
17. Krishna, R., Zhu, Y., Groth, O., Johnson, J., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (May 2017)
19. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and Brain Sciences **40** (2017)
20. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553) (2015)

21. Liu, D., Bober, M., Kittler, J.: Visual semantic information pursuit: A survey. IEEE transactions on pattern analysis and machine intelligence (2019)
22. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep Learning for Generic Object Detection: A Survey. International Journal of Computer Vision **128**(2), 261–318 (Feb 2020)
23. Mancini, M., Karaoguz, H., Ricci, E., Jensfelt, P., Caputo, B.: Knowledge is Never Enough: Towards Web Aided Deep Open World Recognition. In: IEEE ICRA. p. 9543 (May 2019)
24. Marcus, G.: Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631 (2018)
25. Marino, K., Salakhutdinov, R., Gupta, A.: The More You Know: Using Knowledge Graphs for Image Classification. In: Proceedings of IEEE CVPR. pp. 20–28 (Jul 2017)
26. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks **113**, 54–71 (May 2019)
27. Paulius, D., Sun, Y.: A survey of knowledge representation in service robotics. Robotics and Autonomous Systems **118**, 13–30 (2019)
28. Pearl, J.: Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution. In: Proceedings of WSDM 2018. p. 3. ACM (Feb 2018)
29. Serafini, L., Garcez, A.d.: Logic Tensor Networks: Deep Learning and Logical Reasoning from Data and Knowledge. arXiv:1606.04422 [cs] (Jul 2016)
30. Storks, S., Gao, Q., Chai, J.Y.: Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. arXiv preprint arXiv:1904.01172 (2019)
31. Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., van den Hengel, A.: Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding **163**, 21–40 (2017)
32. Yang, K., Russakovsky, O., Deng, J.: Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2051–2060 (2019)
33. Young, J., Kunze, L., Basile, V., Cabrio, E., Hawes, N., Caputo, B.: Semantic web-mining and deep vision for lifelong object discovery. In: Proceedings of ICRA. pp. 2774–2779. IEEE (2017)
34. Zeng, A., Song, S., Yu, K.T., Donlon, E., Hogan, F.R., Bauza, M., Ma, D., Taylor, O., Liu, M., Romo, E., et al.: Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. In: 2018 IEEE ICRA. pp. 1–8. IEEE (2018)