Towards Visually Intelligent Agents (VIA): a Hybrid Approach

Agnese Chiatti
Middle-Late stage PhD Student

@agnese_chiatti
achiatti.github.io

Supervisors: Prof. Enrico Motta, Dr. Enrico Daga

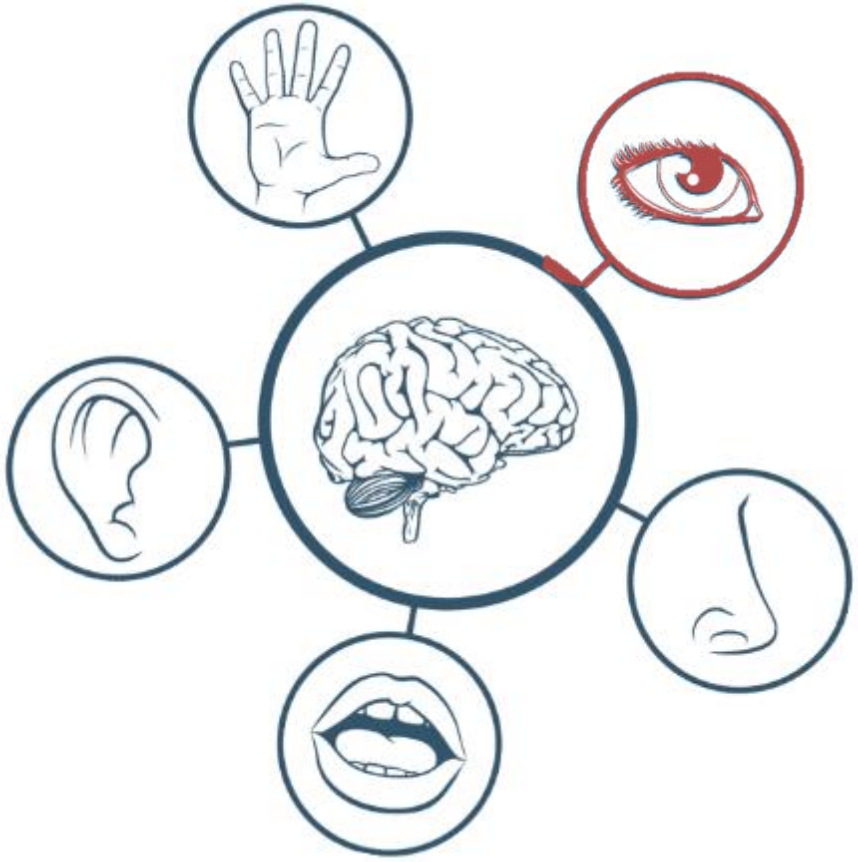
Knowledge Media Institute

The Open University

Service Robotics

HanS:
the Health & Safety
Robot Inspector

Research scope



From **perception**

to **sensemaking**

Visual Intelligence (VI)

"a robot's ability to use its vision system,
reasoning components
and background knowledge
to make sense of the environment."
(Chiatti *et al.*, 2020)

Chiatti, A., Motta, E., Daga, E. (2020) *Towards a Framework for Visual Intelligence in Service Robotics: Epistemic Requirements and Gap Analysis*. In Proceedings of KR2020 – Special Session on KR & Robotics.

Background

ML limitations (Marcus, 2018; Pearl, 2018; Parisi et al., 2019)

Promise of **hybrid methods** (Aditya et al., 2019; Gouidis et al., 2019)

Knowledge-based components **at different levels**

Post-hoc integration: modularity, isolating the contributing components, transparency but also knowledge reliabilty and computational overhead to keep in check

Advances in SW and KE: abundance of resources but which to prioritise? (Daruna et al., 2018)

Research Questions

Hypothesis: *a hybrid approach (ML+knowledge-based) can improve a robot's performance on tasks that require VI, compared to pure ML.*

RQ1: *What are the epistemic requirements of developing VIA?*

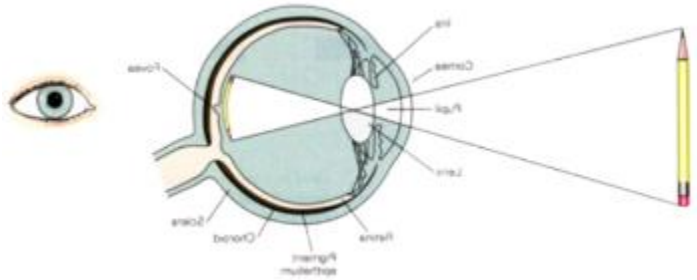
RQ2: *Which epistemic requirements are* ***the most important ones****, in our use-case?*

RQ3: *To what extent do* ***state-of-the-art KBs*** *support VIA?*

RQ4: *To what extent can existing KBs be* ***repurposed*** *to support the requirements highlighted in RQ2?*

RQ5: *To what extent can* ***a concrete architecture*** *which integrates the identified types of reasoners be developed?*

RQ1: Top-down analysis



LEARNING = MODEL BUILDING

NAIVE PHYSICS

Infants can already grasp basic principles of Physics before 6 months of age.

GENERIC 2D VIEWS

The images cast at the back of our eye are 2-dimensional. We construct the 3D mentally.

COMPOSITIONALITY

We are able to segment the different parts composing an object and to disambiguate each object from other ones located nearby.

MOTION VISION

We categorise objects and actions differently based on the perceived movement (or lack thereof).

FAST PERCEPTION

Our visual perception is extremely fast and we can learn to recognize new objects even from the very first exposure.

RQ1: Bottom-up analysis

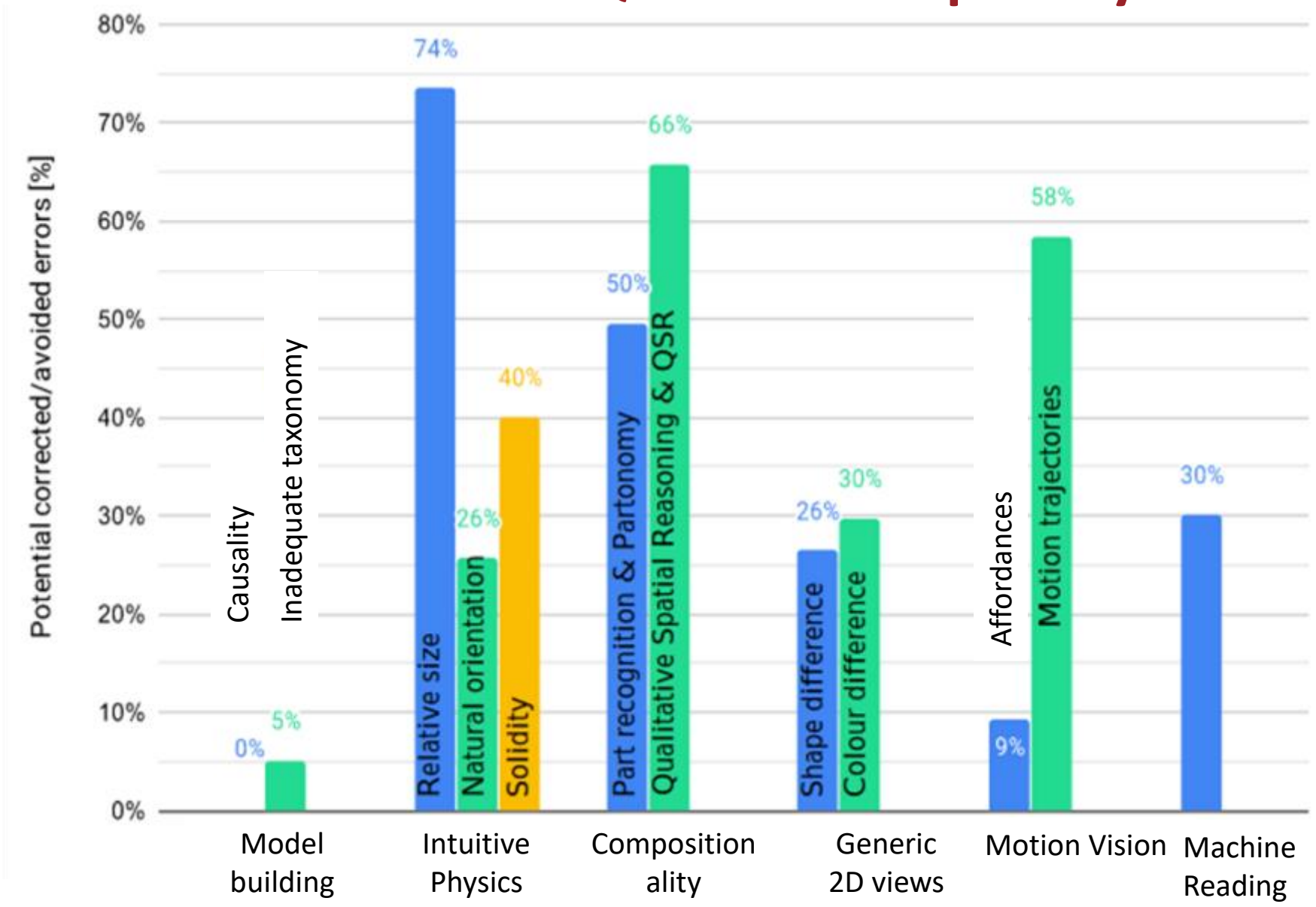
Object recognition pipeline **purely based on ML**
(Chiatti et al., 2020)

Out of 896 test regions, 272 (31.59%) were mis-classified

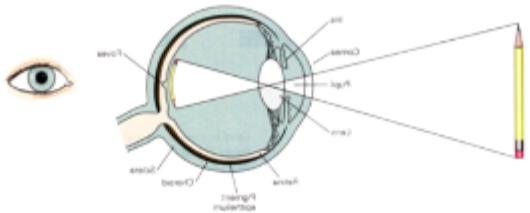
Which capability/-ies or knowledge property/-ies can help?

Ground truth class	Predicted class	Intuitive Physics			Spatial Reason-ing	Machine Reading
		Size	Ori.	Solid.		
Bin	Mug	T	F	F	T	F
Bin	Mug	T	F	F	T	T

RQ2: Bottom-up analysis



Visual Intelligence Framework



GENERIC 2D VIEWS

Use of prototypical 2D shapes for fast object classification

NAIVE PHYSICS

Physical properties of objects (e.g., size, natural orientation, etc.)

COMPOSITIONALITY

Spatial reasoning; part-whole relations; fine-grained segmentation

LEARNING = MODEL BUILDING

MOTION VISION

object tracking and action recognition across temporally ordered frames

FAST PERCEPTION

DL methods provide near real-time object recognition performance on known categories

MACHINE READING

text recognition and understanding from labelled objects and signs

RQ3: KB coverage study

KB	Knowledge Requirements of Visual Intelligence								Accessibility
	Hierarchical Taxonomy (linked to WordNet?)	Cause-effect relations	Intuitive Physics Knowledge	Part-whole relations	QSR	Generic 2D views	Object affordances	Motion tra-jectories	
Unified Verb Index (UVI)	yes	●					●		High
KnowRob/Open-EASE	yes	●	●●	●			●	●	Partial
DBpedia	yes		●	●	●		●		High
Wikidata	yes		●	●	●	●	●		High
Probase	no	●●	●	●	●				Partial
NELL	no		●	●	●		●●		Adequate
ConceptNet	yes	●●	●●	●	●		●●		High
WebChild	yes		●	●	●		●●		High
ATOMIC	no	●					●		High
ASER	no	●●					●		Adequate
Visual Genome (VG)	yes		●●	●●	●●	●	●●		High
ShapeNet/ PartNet	yes		●●	●●		●●			Partial

Chiatti, A., Motta, E., Daga, E. (2020) *Towards a Framework for Visual Intelligence in Service Robotics: Epistemic Requirements and Gap Analysis*. In Proceedings of KR2020 – Special Session on KR & Robotics.

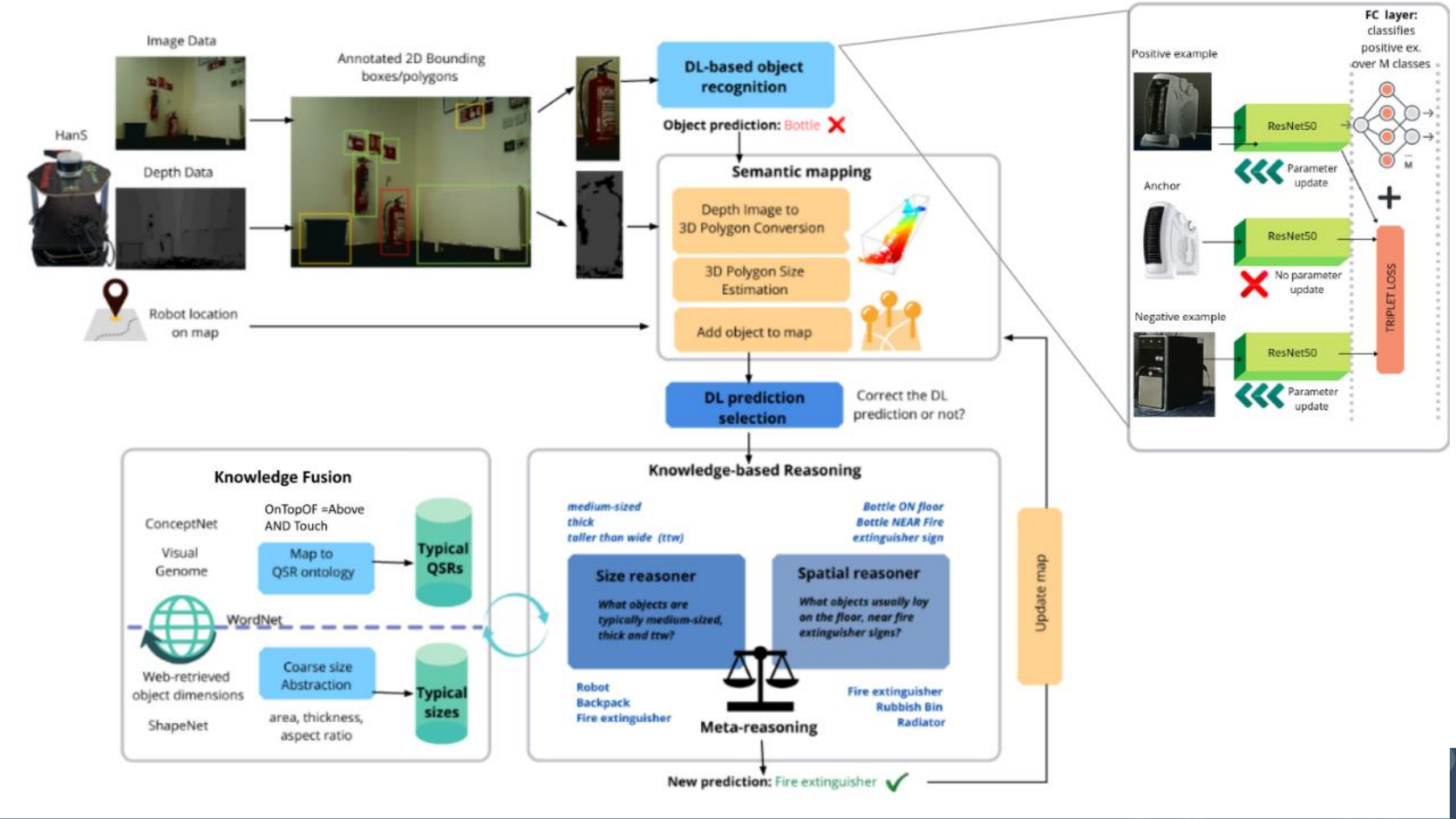


Image Data
Annotated 2D Bounding boxes/polygons
HanS
Depth Data
Robot location on map
DL-based object recognition
Object prediction: Bottle
Semantic mapping
Depth Image to 3D Polygon Conversion
3D Polygon Size Estimation
Add object to map
DL prediction selection
Correct the DL prediction or not?
FC layer: classifies positive ex. over M classes
Positive example
Anchor
Negative example
ResNet50
ResNet50
ResNet50
Parameter update
No parameter update
Parameter update
TRIPLET LOSS
Knowledge Fusion
ConceptNet
Visual Genome
WordNet
ShapeNet
Web-retrieved object dimensions
OnTopOF =Above AND Touch
Map to QSR ontology
Typical QSRs
Coarse size Abstraction
area, thickness, aspect ratio
Typical sizes
Knowledge-based Reasoning
medium-sized
thick
taller than wide (ttw)
Bottle ON floor
Bottle NEAR Fire extinguisher sign
Size reasoner
What objects are typically medium-sized, thick and ttw?
Spatial reasoner
What objects usually lay on the floor, near fire extinguisher signs?
Robot
Backpack
Fire extinguisher
Meta-reasoning
Fire extinguisher
Rubbish Bin
Radiator
New prediction: Fire extinguisher
Update map

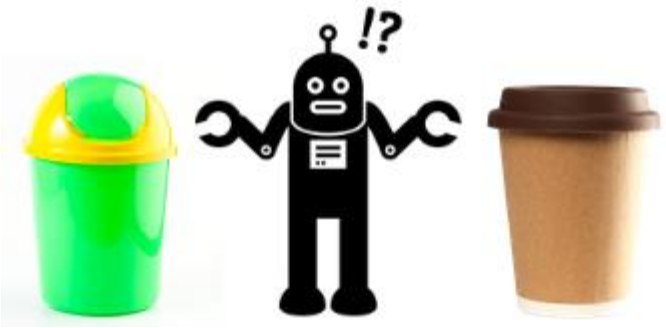
Representing sizes

Mid-level representation, robust to **contour variance**
(Long et al., 2016; Zhu et al., 2014; Bagherinezad et al., 2016; Elazar et al., 2019)

Coarse features but also fine-grained enough for categorisation
→ a **synoptic representation of size**

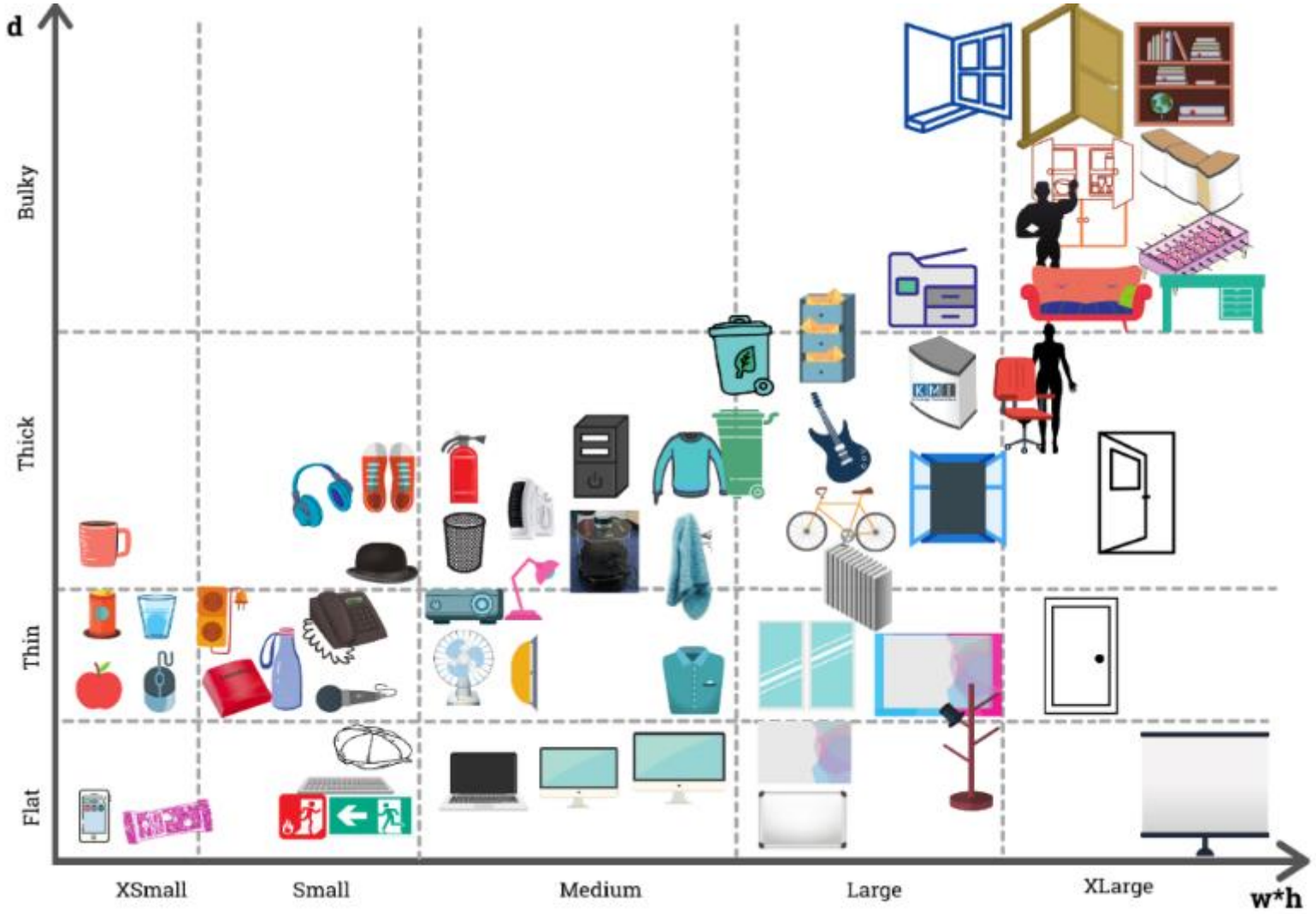
Object **surface area**, **depth** and **Aspect Ratio (AR)**

Automatically abstracted from multiple raw size measurements
(ShapeNet, Amazon, manually-collected)



Chiatti, A., Motta, E., Daga, E., Bardaro, G. (2021) *Fit to Measure: Reasoning about Sizes for Robust Object Recognition*. In Proceedings of the AAAI-MAKE 2021 Spring Symposium.

Proposed Representation



Experimental results: KMi dataset

On a test set of 1414 object regions, annotated w.r.t. 60 reference classes

	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

Experimental results: Amazon 2017 Image Matching dataset

On a test set of 562 images, annotated w.r.t. 61 classes (41 known + 20 novel)

	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

Representing Qualitative Spatial Relations (QSR)

Several different **AI formalisms** (Cohn& Renz., 2008)

Semantic mapping methods & GIS technologies (Kostavelis & Gasteratos, 2015)

Best of both worlds: defining a mapping (Deeken et al., 2018)

But also account for **everyday language use** to describe spatial relations (Landau & Jackendoff, 1993)

Which is also **reflected in KBs** like Visual Genome, ConceptNet, SpatialSense, etc.

Broader impact on **HRI** (Sarthou et al., 2019; Sisbot & Connell, 2019)

Representing Qualitative Spatial Relations (QSR)

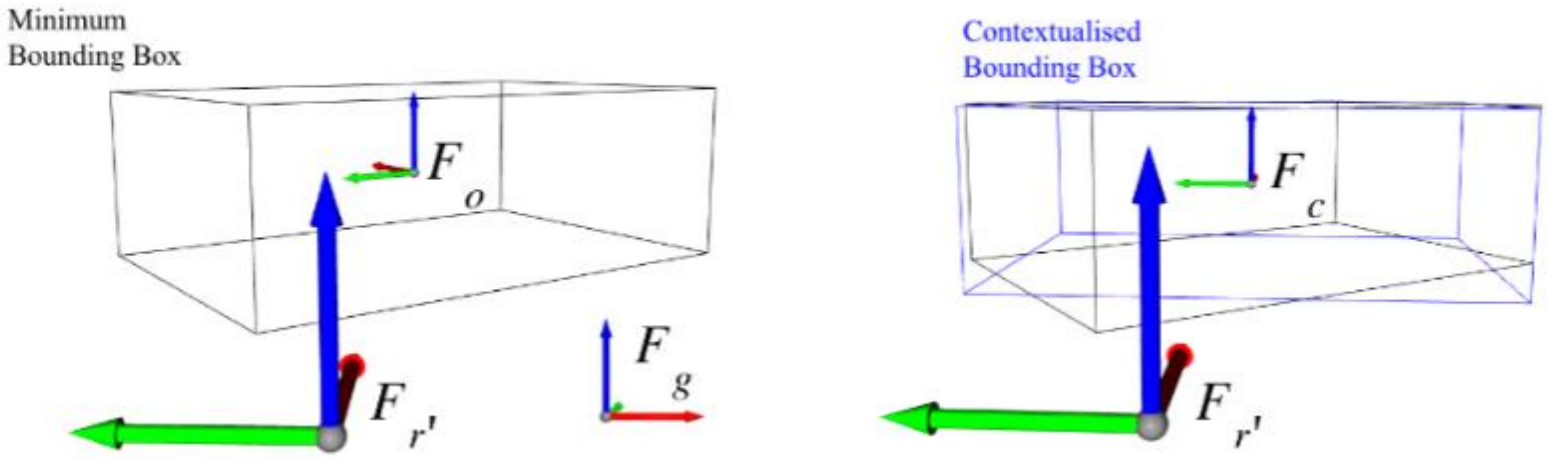
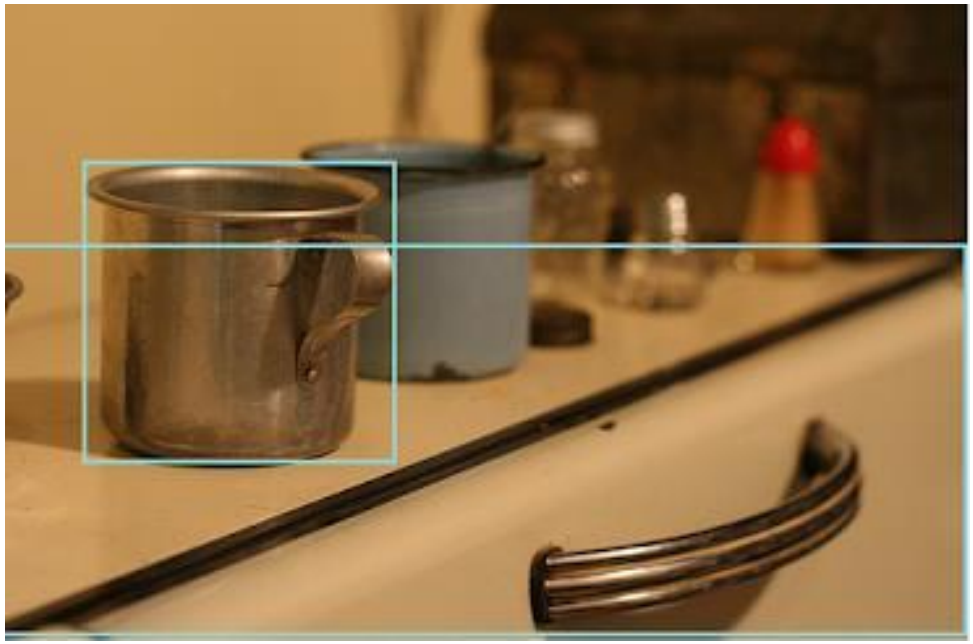


Figure and reference but also dependent on the observer

Chiatti, A., Bardaro, G., Motta, E., Daga, E. (2021) *Commonsense Spatial Reasoning for Visually Intelligent Agents*. https://arxiv.org/abs/2104.00387

Extracting spatial statistics from VG

Different spatial uses of “on” predicate: on top of, leans on, affixed on



Img source: https://visualgenome.org/

Conclusion and next steps

Before we can delegate tasks to robots, we need to enhance their sensemaking capabilities.

Contributed **a framework of requirements for VIA** and **guidelines on priorities**

Preliminary evidence that **knowledge** of the typical size and spatial relations between objects **integrated in post-processing can significantly augment ML**.

Evaluation of the spatial reasoning component in progress

Meta-reasoning: study of cases of agreement/disagreement between reasoners

Completing assessment of utility to support **decision-making scenarios**

ISDS
Intelligent Systems
and Data Science

Thank you!
Q&A

robots.kmi.open.ac.uk

https://github.com/kmi-robots

@agnese_chiatti
@isdsou

Linked with **common sense**

Experiential

Cross-sectional
(Space, Time, Physics,…)

ML limitations