CSK-Detector: Commonsense in object detection

Irina Chernyavsky[1], Aparna S. Varde[1.2] and Simon Razniewski[2]
1. Montclair State University, NJ, USA; 2. Max Planck Institute for Informatics (MPII), Saarbrucken, Germany
(chernyavskyi1 — vardea)@montclair.edu, (avarde — srazniew)@mpi-inf.mpg.de

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10020915

***Abstract*—We propose an approach *CSK-Detector* for object detection and image categorization, well-suited for big data, by transferring commonsense knowledge from a knowledge base, augmented with premises and quantifiers. It is implemented for domestic robotics, especially with the motivation that next-generation and multipurpose domestic robots should be able to seamlessly discern environments for specific tasks without prior annotation of excessive images. CSK-Detector is evaluated on real data, yielding better results than deep learning without commonsense, while also providing an explainable approach. It broadly impacts human-robot collaboration and smart living.**

Keywords: Big data, commonsense knowledge, domestic robotics, explainable AI, image categorization, smart living

I. Introduction

Motivation and Problem Definition: Advances in artificial intelligence enable developing computer vision models to classify images and videos. Although this aids robotics, misclassification on real-world images is yet a problem. As humans, we have subtle knowledge of concepts, properties and relationships, i,e. *commonsense knowledge (CSK)*. It helps us intuitively categorize images, even at first sight. Robots do not have natural CSK, and thus classify images based on prior training. Many algorithms thrive on neural networks and deep learning, i.e. a black-box without CSK, possibly erring in first-time tasks, requiring huge training data, and lacking explainability. This motivates object detection and image categorization infused with CSK to address the following.

- Relevance of the detected object to a category
- Explainability of the categorization algorithm
- Easy adaptation to other autonomous systems
- Automated dataset preparation for image learning

The solution should not require annotating an overwhelmingly large number of full images to create prior labeled training data. Deep learning models require such huge labeled datasets; more the data, better the learning. If image annotations have errors, the learning can be flawed. Furthermore, they use general domains, hard to apply directly to specific tasks.

Related Work: Some works classify house space for mobile robots via topology and semantics, fusing multiple visual cues & laser range data [1]. Others use geometry of spaces through CNN objects [2], probabilistic hierarchical models [3], or local/global discriminating data [4]. Such approaches often use deep learning with excessive prior annotation of each full image. They lack commonsense and explainability. Commonsense based approaches can produce excellent results in various machine learning tasks [5], [6], [7]. [8].

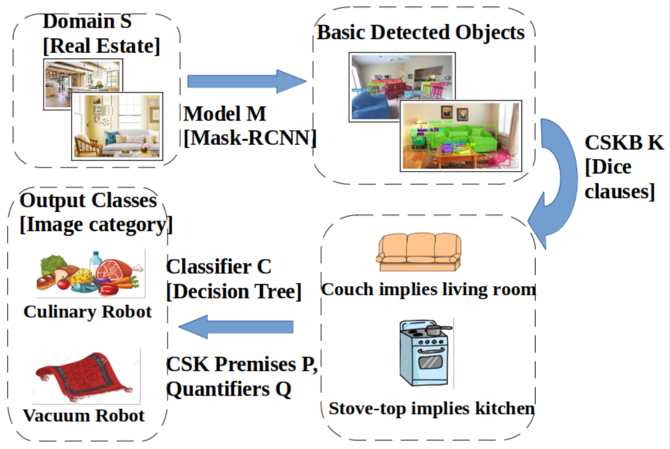

Fig. 1. Overview of CSK-Detector approach (executed in domestic robotics)

Contributions: We propose an approach called *CSK-Detector: **C**ommon **S**ense **K**nowledge based object* **Detector**, with the following contributions.

1) It does not need prior annotation of each full image
2) It uses a visualizable, explainable classifier
3) It imbibes CSK to reduce misclassification
4) It can generalize/specialize to other autonomous tasks

Fig. 1 depicts an overview of CSK-Detector, as per its implementation in domestic robotics. Its details are described next.

II. Proposed Approach: CSK-Detector

Our CSK-Detector approach has the following parts.

1) Source domain S with basic detection model M
2) CSKB K to find object relationships in images
3) Explainable classifier C such as a decision tree
4) Simple CSK premises P with quantifiers Q

It outputs image category based on relevant object detection.

1. For S, we use *real estate*, employing M as *Mask-RCNN*. It resizes images, and detects *basic objects*, e.g. couch, TV etc. *Note that there are just a few basic objects in homes **(100s)**, vs. many growing real-world images **(millions)***.

2. For K, we harness the multifaceted *Dice* CSKB (Commonsense Knowledge Base) with joint reasoning on sets of interrelated statements. It has 4 facets: *plausibility, typicality, remarkability, salience* [9]; each has a score, relying on soft constraints, with logical clauses, e.g. Eq. 1, 2.

$$Pl(s_1,p) \wedge Rl(s_1,s_2) \wedge \neg Pl(s_2,p) Rm(s_1,p) \tag{1}$$

$$Ty(s,p) \wedge Rm(s,p) Sa(s,p) \tag{2}$$

Here, Pl means *plausible*, Rl is *related*, Rm is *remarkable*, Ty is *typical*, Sa is salient, s, s_1, s_2 are semantic entities or concepts (e.g. rooms, chores), and p is a property (e.g. objects in rooms). If it is *plausible* to find a *bed* in a *bedroom* and *not plausible* to find it in a related entity *kitchen*, a *bed* is *remarkable* for a *bedroom* and gets a high remarkability score there. Using scores for each facet, aggregated scores are learned via a regression model in Dice. We reuse these, calling them A scores in CSK-Detector. Each detected object per image receives an A score. It is used to assign classes on a uniform scale of 4: [0.0–0.24] : no, [0.25–0.49] : low, [0.5–0.74] : $medium$, [0.75–1.0] : $high$, e.g. if *TV* has $A = 0.8$ for *bedroom*, it has a high chance of being there. This logic extends to object-combinations via average aggregate scores, calculated as A_{avg} for n distinct objects as in Eq.3.

$$A_{avg} = (A_1 + A_2 + ... + A_n)/n \tag{3}$$

For instance, if *spices*, *sink* are detected in an image with A scores 0.91, 0.58 respectively for the concept *kitchen*, then $A_{avg} = 0.75$ (high chance that the image is a kitchen).

3. Information is fed to C, a decision tree classifier by creating a dataset of all the scores and an intermediate class (e.g. room=kitchen: no, low, medium, high). This forms training data for C to learn a hypothesis H.

4. We propose commonsense premises P in CSK-Detector as per concepts in K (Dice). These, along with the learned hypothesis H help to gauge the final image categories (e.g. culinary). Premises P entail positive and negative clauses for added emphasis. Examples are in Eq. 4, 5.

$$\textit{Room} = \textit{Kitchen} \Rightarrow \textit{Class} = \textit{Culinary} \tag{4}$$

$$\textit{Chore} = \textit{WashClothes} \Rightarrow \textit{Class} = \neg\textit{Vacuum} \tag{5}$$

We propose quantifiers Q for objects. By our own commonsense, it is not feasible to have more than 2 ovens in a kitchen, so if that occurs in an image, it can be something else, e.g. showroom. Various objects have different Q values with upper limits L. We pre-define these using basic CSK and store them for ∼100 potential objects in S (e.g. oven, bed). They are used by CSK-Detector for filtering, e.g. *"o1=oven, q1=3: class=none"* (*q1* is quantifier for object *o1*).

Given all this discussion, the algorithm for CSK-Detector appears as Algorithm 1, based on our execution.

Algorithm 1: CSK-Detector Processing

```
Input: Images in $S$, Quantifier Limit $L$ per object
1. Pass images through Mask-RCNN
2. Return basic detected objects from Mask-RCNN
3. For each object $O$, find quantifier $Q$
4. If $Q > L$, return "none: not relevant"
5. Else calculate multifaceted scores: Dice & Eq.3
6. Build training data: objects, scores, intermediate class
7. Pass training data to Decision Tree to learn hypothesis $H$
8. Use $H$, CSK-Premises $P$ to trace final Image Category
Output: Image Category (Culinary etc.)
```

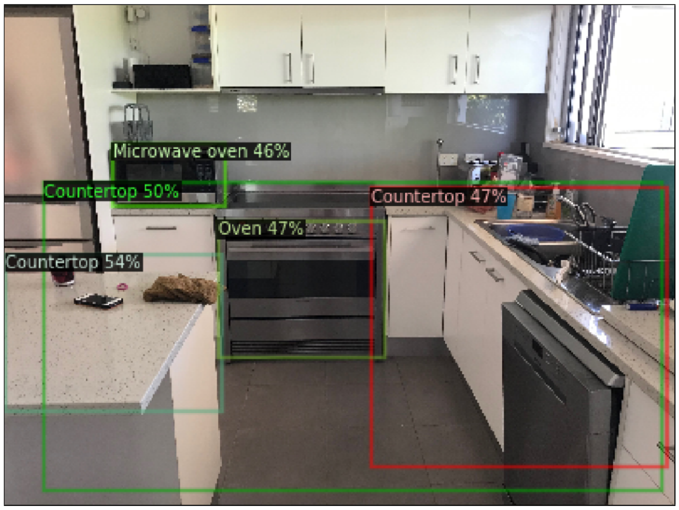

Fig. 2. Examples of basic detected objects in an image

Object	Limit
refrigerator	2
microwave	2
sink	2
blender	2
TV	1
dishwasher	1
fruits	15
utensil	10
cabinet	4
coffee maker	2
rice cooker	1
kettle	3
oven	2
table	2

TABLE I
EXAMPLE OF QUANTIFIER UPPER LIMITS FOR A FEW OBJECTS

III. Implementation in Domestic Robotics

CSK-Detector is implemented in the context of domestic robotics in this paper. We obtain images online from the real estate domain. These images are resized to the height, width of 430 pixels, and 640 pixels respectively. They are preprocessed and passed into Detectron2 (PyTorch-based modular object detection library), which uses Mask-RCNN for object detection. It detects basic objects, e.g. refrigerator, couch, table, chair, sofa, countertop (see Fig.2). Each object is assigned a cutoff as a quantifier upper limit L based on commonsense knowledge, e.g. a few houses may have 2 refrigerators but it is odd to see 3 refrigerators in a household. Hence, 3 is a quantifier limit L for refrigerator w.r.t kitchen (so if 3+ refrigerators are detected in the image, it is considered irrelevant to a household, and hence to domestic robotics, e.g. it may be an image of a retail store showroom such as Home Depot). See Table I for examples of L for some basic objects.

After selecting relevant images based on quantifiers and limits, detected objects are passed through relational clauses derived from the multifaceted CSKB *Dice* [9]. Its statements are simple, e.g. "refrigerator in a kitchen", "couch in a living room”. They carry weighted soft constraints for reasoning on 4 facets: plausibility, typicality, remarkability, salience. Plausibility indicates if the statement makes sense w.r.t. a concept and its properties, e.g. does it make sense to see a bathtub in the kitchen? (No). Typicality means that a property holds for most instances of the concept, e.g. utensils in the kitchen. Remarkability is a property that distinguishes it from its siblings (i.e. other similar concepts), e.g. spices are remarkable in the kitchen because it is odd to find them in other rooms of a house. Salience is a property truly characteristic of a concept (salient feature), so it must be remarkable & typical.

Each facet carries a score. Scores are fed into a regression model in Dice to learn an aggregate score. Aggregate scores derived from Dice, denoted as A scores in CSK-Detector, are used as guiding scores to assign classes. For instance, given a concept (room) “kitchen” and a property (object) “sink”, there can be many statements in Dice about both of them: “sink is found in kitchen”, “sink is located in kitchen”, “sink is near kitchen” and so on. They have their respective aggregate scores based on plausibility, typicality etc. Thus, A scores in CSK-Detector are the average aggregate scores of all such statements. Likewise, A scores for combinations of 2 or more objects (e.g. sink, oven) are calculated as their average A_{avg} as in Eq. 3, where n is the number of distinct objects.

For instance, w.r.t. kitchen, the A score for “spices” can be combined with that for “sink” to obtain an average A score for both, and hence classify an image. As per our own commonsense knowledge, we can gauge that if an image has spices and a sink, it is likely to be a kitchen, but if it has a bathtub and a sink, it is likely to be a bathroom. Such combinations can be addressed via average A scores.

Hence, these combinations help to better distinguish concepts in images and thereby classify them into intermediate classes (kitchen etc.) which are then fed to the decision tree classifier. See Table II for a partial snapshot of the data used for learning a hypothesis H in the classifier. (Note: Here *Pl* is Plausible, *Ty* is Typical, *Rm* is Remarkable, *Sa* is salience, *A* denotes aggregate i.e. A score, and *C:kitchen* stands for class=“kitchen”)

Object	*Pl*	*Ty*	*Rm*	*Sa*	*A*	*C:kitchen*
refrigerator	0.44	0.03	0.51	0.57	0.58	medium
microwave	0.23	0.48	0.01	0.16	0.75	high
countertop	0.71	0.48	0.01	0.13	0.58	medium
sink	0.38	0.23	0.73	0.62	0.58	medium
cabinet	0.55	0.15	0.98	0.7	0.96	high
fruits	0.23	0.48	0.01	0.22	0.46	medium
spices	0.35	0.02	0.99	0.29	0.91	high
pot	0.64	0.98	0.97	0.75	0.96	high
sofa	0.16	0.41	0.07	0.18	0.25	low
bed	0.23	0.03	0.21	0.24	0.08	low

TABLE II
PARTIAL SAMPLE OF DATA FED INTO THE DECISION TREE CLASSIFIER

As mentioned earlier, these intermediate classes are uniformly assigned on a scale of four $(25, 50, 75, 100)$, based on the A score for each object (property) as per that concept. Thus, if ”refrigerator” has A score as 0.58 “to be in a kitchen”, it indicates a medium chance of an image being a “kitchen” based on that object alone. If that along with another object gives an average aggregate score of 0.75 or higher, the image with both these objects can have a higher chance of being a kitchen. The same logic extends to other combinations of 3 objects or more.

The dataset is created based on the 4 facets scores (plausibility, typicality, remarkability, salience), the aggregate score and the intermediate class. Once these are fed to the decision tree classifier, it learns a hypothesis H. This serves as the basis to categorize images into final classes (culinary / laundry etc.), using our commonsense premises P. For example, if the intermediate class is a “kitchen” (medium / high), the final class or image category is assigned as “culinary”, based on the commonsense premise in Eq. 4.

Listing 1: Sample Commonsense Premises

$Chore = WashClothesClass = Laundry$
$Chore = FoldingClass = Laundry$
$Chore = CookingClass = Culinary$
$Chore = WashVegetablesClass = Culinary$
$Chore = FryingClass = Culinary$
$Chore = ChoppingClass = Culinary$
$Chore = MincingClass = Culinary$
$Chore = MopClass = Vacuum$
$Chore = SweepClass = Vacuum$
$Chore = WipeDownClass = Vacuum$
$Room = BedroomClass = \neg Culinary$
$Room = LivingRoomClass = \neg Culinary$
$Room = BathroomClass = \neg Culinary$
$Chore = CookingClass = \neg Vacuum$
$Chore = FryingClass = \neg Laundry$

Likewise, we list a few more commonsense premises in Listing 1. These are used to detect final image categories (Algorithm 1). Hence, CSK-Detector can produce its output.

IV. EXPERIMENTAL EVALUATION

We summarize our evaluation of CSK-Detector as per domestic robotics. In our experiments, datasets have 2140 *real estate* images from Kaggle (indoor, outdoor). Abstract and retail store images are added, e.g. icons of smart cities, Home Depot showrooms etc. Ground truth is mainly based on existing image captions. Training data for CSK-Detector (classifier C) is as explained the previous section; n-fold cross-validation is used for testing (n=4,10). It gives single-class outputs (each image is in its most likely category). We compare CSK-Detector with other approaches, e.g. VGG16 [10], AlexNet [11]. Evaluation is summarized in Table III.

Approach	Accuracy
VGG16	88.54%
ResNet	91.54%
EfficientNetB5	92.01%
Xception	62.38%
AlexNet	73.72%
CSK -Detector	91.25%

TABLE III
COMPARATIVE EVALUATION OF CSK-DETECTOR WITH % ACCURACY

It is observed that CSK-Detector performs well in comparative studies (with the same dataset for all approaches), While evaluation is performed with a sample of real data, we have designed CSK-Detector such that it can easily scale well to larger datasets. We briefly discuss our experimental results, highlight their merits, and point out the scope for improvement. Examples of correctly and incorrectly classified images from CSK-Detector appear in Figs 4, 4, 5. For simplicity, we show an example of an image in each category (culinary/vacuum/laundry) with misclassification & correct classification, respectively.

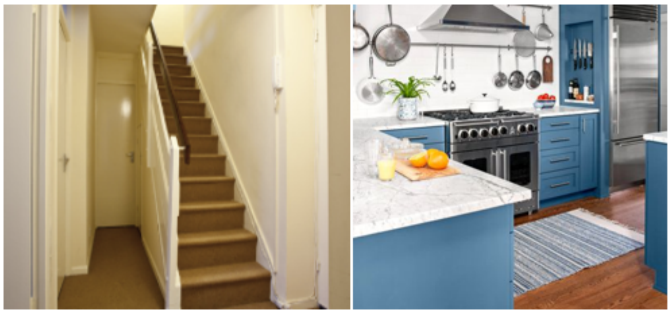

Fig. 3. Misclassified (left) and correctly classified (right): Culinary category

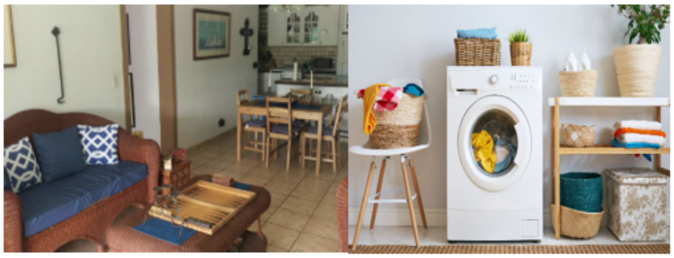

Fig. 4. Misclassified (left) and correctly classified (right): Laundry category

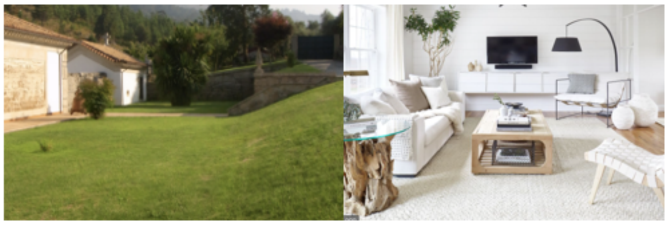

Fig. 5. Misclassified (left) and correctly classified (right): Vacuum category

As observed in these examples, the misclassification can happen due to lack of commonsense knowledge derived from the logical clauses within the concerned CSKB utilized or due to incorrect basic object detection. For example, in Fig. 5, the misclassified image looks like a carpet (hence is detected thus by the Mask-RCNN), and is mistaken for the carpet by the classifier within CSK-Detector, hence considering the image to be in the vacuum category. In Fig. 4 the misclassification results from a smaller dataset for the laundry category, as found in the CSKB Dice. Hence, inclusion of more logical clauses, and increasing the CSKB size as well as versatility can provide a better classification. This motivates development of domain-specific CSKBs, e.g. a domestic robotics CSKB, so as to enhance image classification and other activities in AI and robotics. Additionally, we can harness spatial collocations via related systems [12] to enhance basic object detection, reduce misclassification, and improve CSK-Detector performance.

V. Conclusions and Future Work

This paper proposes an approach *CSK-Detector* to detect objects and categorize images for automated learning, and is executed in domestic robotics. It is an explainable approach based on transferring commonsense knowledge, with premises and quantifiers; easy to fathom and modify for specific purposes / generalize to other autonomous systems. CSK-Detector has similar or higher accuracy vs. approaches with deep learning black-box models. Our work opens further research: (1) refine methods to raise accuracy; (2) add images, target classes, multiclass outputs, and spatial collocations; (3) extend to other domains. CSK-Detector can potentially aid human-robot collaboration [13] as well as next-generation multipurpose robots [14] for smart living.

Acknowledgments

A. Varde has *NSF grant 2117308* in robotics. She is a visiting researcher at *MPII Germany* (ongoing from sabbatical).

References

[1] A. Pronobis, O. Martinez Mozos, B. Caputo, and P. Jensfelt, "Multi-modal semantic place classification," *The International Journal of Robotics Research*, vol. 29, no. 2-3, pp. 298–320, 2010.

[2] D. Chaves, J.-R. Ruiz-Sarmiento, N. Petkov, and J. Gonzalez-Jimenez, "Integration of CNN into a robotic architecture to build semantic maps of indoor environments," in *Intl. Conf. on ANN*, pp. 313–324, 2019.

[3] P. Espinace Ronda, T. Kollar, A. Soto, and N. Roy, "Indoor scene recognition through object detection," in *ICRA*, pp. 1406–1413, 2010.

[4] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *IEEE CVPR*, pp. 413–420, 2009.

[5] N. Tandon, A. S. Varde, and G. de Melo, "Commonsense knowledge in machine intelligence," *SIGMOD Record*, vol. 46, pp. 49–52, 2017.

[6] A. Pandey, M. Puri, and A. Varde, "Object detection with neural models, deep learning and common sense to aid smart mobility," in *2018 IEEE 30th international conference on tools with artificial intelligence (ICTAI)*, pp. 859–863, IEEE, 2018.

[7] A. Garg, N. Tandon, and A. S. Varde, "I am guessing you can't recognize this: generating adversarial images for object detection using spatial commonsense (student abstract)," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 13789–13790, 2020.

[8] S. Razniewski, N. Tandon, and A. S. Varde, "Information to wisdom: Commonsense knowledge extraction and compilation," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 1143–1146, 2021.

[9] Y. Chalier, S. Razniewski, and G. Weikum, "Joint reasoning for multi-faceted commonsense knowledge," in *AKBC*, 2019.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[12] A. Garg, N. Tandon, and A. Varde, "CSK-Sniffer: Commonsense Knowledge for Sniffing Object Detection Errors," in *ACM EDBT BigVis*, 2022.

[13] C. J. Conti, A. S. Varde, and W. Wang, "Human-robot collaboration with commonsense reasoning in smart manufacturing contexts," *IEEE Transactions on Automation Science and Engineering*, 2022.

[14] Mira-Robotics-Japan, "Ugo - the multi-purpose household robot of the future." www.dw.com/en/ugo-the-multi-purpose-household-robot-of-the-future/video-55585607, 2021.