Exploring Task-agnostic, ShapeNet-based Object Recognition for Mobile Robots

Agnese Chiatti
Knowledge Media Institute
The Open University
United Kingdom
agnese.chiatti@open.ac.uk

Gianluca Bardaro
Knowledge Media Institute
The Open University
United Kingdom
gianluca.bardaro@open.ac.uk

Emanuele Bastianelli
The Interaction Lab
Heriot-Watt University
United Kingdom
emanuele.bastianelli@hw.ac.uk

Ilaria Tiddi
Faculty of Computer Science
Vrije Universitet Amsterdam
The Netherlands
i.tiddi@vu.nl

Prasenjit Mitra
Information Sciences and Technology
The Pennsylvania State University
Pennsylvania, USA
pmitra@ist.psu.edu

Enrico Motta
Knowledge Media Institute
The Open University
United Kingdom
enrico.motta@open.ac.uk

ABSTRACT

This position paper presents an attempt to improve the scalability of existing object recognition methods, which largely rely on supervision and imply a huge availability of manually-labelled data points. Moreover, in the context of mobile robotics, data sets and experimental settings are often handcrafted based on the specific task the object recognition is aimed at, e.g. object grasping. In this work, we argue instead that publicly available open data such as ShapeNet [8] can be used for object classification first, and then to link objects to their related concepts, leading to task-agnostic knowledge acquisition practices. To this aim, we evaluated five pipelines for object recognition, where target classes were all entities collected from ShapeNet and matching was based on: (i) shape-only features, (ii) RGB histogram comparison, (iii) a combination of shape and colour matching, (iv) image feature descriptors, and (v) inexact, normalised cross-correlation, resembling the Deep, Siamese-like NN architecture of [31]. We discussed the relative impact of shape-derived and colour-derived features, as well as suitability of all tested solutions for future application to real-life use cases.

1 INTRODUCTION

Autonomous sensemaking under rapidly-evolving and uncertain circumstances goes beyond building intelligent and knowledge-based systems, requiring mobile systems that are not only able to reason on their surroundings, but also to readily adapt to their *context*. Context is, first and foremost, bound to the physical objects spread around the observed space, all belonging to different categories, and holding static or dynamic qualities, based on their evolution over time. Scalable and adaptable object recognition through mobile robots is then of crucial importance for successful knowledge acquisition and mapping from rapidly-evolving environments. In fact, accurate object recognition is the essential prerequisite to a number of applications in Robotics, including but not limited to: health and safety monitoring [2], retrieving entities across space through human instructions provided in natural language [18], pre-emptive obstacle removal, particularly in the context of elderly care [1, 20], door-to-door garbage collection in Smart Cities [13]. In this scenario, the ability to generalise across different domains by learning features independently from the end goal, e.g., grasping or mapping, can allow agents to flexibly switch between different tasks and capability sets [32].

State-of-the-art supervised approaches in object recognition from natural scenes [23–25] imply the availability of large collections of labelled examples and lack flexibility, when applied on unseen classes and mutable environments. On the other hand, fully unsupervised approaches can provide exploratory insights and guidelines that, however, require significant further tuning and error analysis. These evidences provide much incentive to explore alternative semi-supervised approaches, to balance out the accuracy and precision of the recognition process with the scalability of the achieved solution. Besides, the recent availability of open, multimodal common sense knowledge [8, 10, 30], has expanded the opportunities to further refine, ground and enrich the extracted object entities.

To form a task-agnostic image representation that enables object recognition under varying classes and conditions, different features can come into play. For instance, chairs and plants can be discriminated from one another, in principle, thanks to their shape alone. However, coat hangers could be mistaken for plants, if colours were not taken into account. Hence, the contribution of shape and colour to the resulting classification needs careful assessment, before applying Neural Net-based methods that can produce less interpretable results, with respect to feature importance. Further, relying on ShapetNet-derived models [8] for similarity matching provides readily available data, already segmented and labelled, while also linking object entities with a set of related concepts, for future knowledge grounding.

Based on these premises, we interrogated on: (i) the relative impact of shape and colour features on the overall object recognition performance, when the presence of errors propagated from prior segmentation faults is minimised, (ii) the scalability and performance of Siamese-like approaches already proven successful for person re-identification [31], when applied to ShapeNet based object recognition instead. To tackle these questions, we designed five pipelines, as the starting point to weigh up further application on images collected on a mobile robot.

In this paper, we present our main contributions, with respect to:

- Assessing the relative importance of shape-derived, colour-derived and hybrid features when similarity-based matching is applied against entities from ShapeNet.
- Evaluating the adequacy of feature descriptors in providing a less expensive and more general object representation, when applied to ShapeNet 2D views.
- Learning object similarity through inexact matching and a CNN-based architectures that shares weights in modelling the two input images, in a Siamese fashion, following an approach that has only be applied to person re-identification across successive frames, but not for task-agnostic object recognition.

For all of the above, we conclude with discussing the obtained results and the emerged challenges, which will inform future improvements in this work. All described data, implemented code and pre-trained models are available at our Github repository[1].

2 BACKGROUND AND MOTIVATION

Recent advances in object recognition methods, such as YOLO [23, 24] or Faster R-CNN [25], have significantly improved performance on predetermined sets of object classes, thanks to expensive *ad hoc* training on manually-labelled data.

The costs and lack of flexibility associated with said solutions, especially when dealing with autonomous agents, have fuelled efforts in designing a number of unsupervised and semi-supervised methods, requiring limited labelled data points and ensuring more abstract and general data representations [4, 14, 33]. Along the same lines, recent efforts have emphasised the need for autonomous agents to recognise cross-domain objects and react flexibly to rapidly-evolving contexts [35]. Addressing similar concerns but from a different angle, other proposed methods [5] have used Generative Adversarial Learning on pre-trained Deep Neural Nets, to foster adaptability to new domains.

On the other hand, the reduced explainability of results obtained through Deep Neural Network based methods [11], suggests that a more careful analysis of the contributing features should be combined with "black-box" learning settings, and can benefit all stages of the knowledge discovery process [26]. Furthermore, identifying the most prominent descriptors has the potential to provide better insights on which modules to fine tune when optimising the solution, in terms of both performance and computational costs. As a result, more scalable solutions also represent a more suitable alternative for mobile robot on-board installation [2, 32]. Therefore, these strategies can ultimately ensure a tradeoff between the more expensive and constrained supervised approaches and the more challenging fully-unsupervised scenarios, where objects are autonomously recognised, e.g., based on the dynamics of their environment [12, 17].

As unsupervised and semi-supervised approaches grow in number and become more established in the context of scene segmentation, object classification and object grasping, knowledge acquisition processes applied upfront, for mapping the robot environment within classes carrying semantic meaning (i.e., semantic mapping [22]) mainly rely on handcrafted knowledge bases or are often based on ARTags, to control for the complexity of autonomous object recognition and rather focus on spatial reasoning and rule implementation [2].

To the best of our knowledge, there has been no prior work, in the related literature, evaluating possibilities to extract general features through ShapeNet-based [8] similarity matching, for the purpose of acquiring task-agnostic knowledge. Further, relative impacts are here evaluated by isolating the classification problem from the additional noise carried over from the object segmentation routines. Thus, the integration of ShapeNet in the proposed workflow is not only motivated by the availability of pre-segmented and pre-labelled data that comes with it, but also by the existing link between objects and their related concepts, enabling future knowledge grounding. Already used for learning object intrinsics [29] and for 3D scene understanding [33], models in ShapeNet have in fact never been used to assess the relative importance of colour and shape derived features when classifying objects, nor have they been combined with CNN-based inexact matching methods. Inspired the application of Siamese-like Networks for person re-identification [31, 34], we seek to test whether similar methods can be applied to Shapenet-based object recognition, to ultimately test their ability to scale towards more diverse classes.

3 EXPERIMENTAL SETUP

3.1 Data preparation

The experiments described in this paper involve different combinations of two main datasets. We focused on natural scenes from the NYUDepth V2 collection [21], already annotated and segmented, and reference 2D models derived from the ShapeNet dataset.

NYUSet NYUDepth V2 [21] comprises of 1449 densely labeled pairs of aligned RGB and depth images and is provided with a MatLab Toolbox for basic data retrieval and manipulation. We implemented our own Matlab script - extending the provided methods for segmented entities extraction - in order to mask out each labelled region belonging to one of the target object classes and store them as separate RGB frames. To reduce cross-class imbalances, we further down-sampled the chair examples available to 1000 instances (see also Table 1).

ShapeNetSet. ShapeNet is a large-scale collection of richly annotated 3D models [8], organised into two subsets: (i) ShapeNetCore, covering 55 object classes with about 51,300 unique 3D models, and (ii) ShapeNetSem, consisting of 12,000 more densely-annotated models across 270 categories. For a number of 3D models, 2D views of the object surfaces are available as well. Further, ShapeNet object annotation is based on synsets, i.e., sets of synonyms defined according to the WordNet lexical database [19], and is linked with the ImageNet set as well [10].

We first selected a subset of models, i.e., two for each of the ten object classes of interest. We will refer to this subset as ShapeNetSet1, or SNS1, in the remainder of this paper. Specifically, for most classes, four 2D views of the selected model were collected, or manually-derived by rotating an existing view, when not available. Less window and door examples were included, representing rotation-invariant models, whereas objects that were either more

[1] https://github.com/kmi-robots/semantic-map-object-recognition

complex in nature or more highly-represented and diversified in the NYUset, such as chairs and bottles, were slightly oversampled (see Table 1). Further, we selected a second, larger, subset (ShapeNetSet2, or SNS2) spread across the same object classes, with ten 2D views for each target category. Class-wise cardinalities are outlined in Table 1.

Table 1: Dataset statistics.

Object	ShapeNetSet1	ShapeNetSet2	NYUSet
Chair	14	10	1000
Bottle	12	10	920
Paper	8	10	790
Book	8	10	760
Table	8	10	726
Box	8	10	637
Window	6	10	617
Door	4	10	511
Sofa	8	10	495
Lamp	6	10	478
Total	**82**	**100**	**6,934**

3.2 Shape and Color Feature Matching

To tackle the first question on relative importance of colour and shape features in recognising a specific class of objects, we conducted a first exploratory analysis on the NYUSet, i.e., we evaluated feature matching-based classification methods alone, leaving potential error-propagation from segmentation faults out of the picture. On a similar note, since the segmented regions from the NYUset were extracted through a black mask, while 2D views from ShapeNet lay on a white background, the marginal noise surrounding both the input objects to classify and the reference views to match against had to be reduced. To achieve this, we (i) first converted to grayscale, (ii) applied global binary thresholding (or its inverse, depending on whether the input background was black or white respectively), (iii) contour detection on cascade, and (iv) cropped the original RGB image to the contour of largest area.

Table 2: Cumulative (cross-class) accuracy under comparison, for all configurations in the exploratory trials and for two data sets: (i) images in the NYUset matched against ShapeNetSet1 (SNS1), (ii) views in ShapeNetSet1 (SNS1) matched against ShapeNetSet2 (SNS2).

Approach	Dataset	
	NYU v. SNS1	SNS1 v. SNS2
Baseline	0.10787	0.10
Shape only L1	0.14350	0.18
Shape only L2	0.14537	0.12
Shape only L3	0.15835	0.19
Color only Correlation	0.15965	0.28
Color only Chi-square	0.14537	0.10
Color only Intersection	0.18777	0.29
Color only Hellinger	0.20637	0.32
Shape+Color (weighted sum)	0.20637	0.32
Shape+Color (micro-avg)	0.16945	0.28
Shape+Color (macro-avg)	0.16513	0.22

We then framed the classification task as follows: a set of K Shapenet models, M_c, is defined for $c = 1, \ldots, N$ object classes of interest (i.e., $N = 10$ in this case). Let V_i be the set of 2D views available for each model $m_i \in M_c$, with $i = 1, .., K$. Each input object to classify is thus matched against each single view $v_j \in V_i$, for all K models, and for all N classes. The m_i determining the predicted label is then the argument optimising either a certain similarity or distance function, based on the following approaches.

Shape-only matching. Contours extracted from input samples were matched through the OpenCV built-in similarity function based on Hu moments [15], i.e. moments invariant to translation, rotation and scale. We tested three different variants of this methods, with distance metric between image moments set to be the L1, L2, or L3 norm respectively.

Colour-only matching comparing the RGB histograms of the input image pairs. Similarly to the previous case, we relied on the OpenCV library and tested different comparison metrics, namely Correlation, Chi-square, Intersection and Hellinger distance.

Hybrid matching. The colour-only and shape-only similarity scores obtained in the previous steps were further combined, using three different objective functions. In all hybrid configurations, the selected ShapeNet model m_i was defined as:

$$m_i = \arg\min \Theta \tag{1}$$

Let S and C be the scores obtained with shape-only and colour-only matching when matching all views against each input image, with α and β being their relative weights. Then, the weighted sum of scores is defined as:

$$\theta = \alpha S + \beta C \tag{2}$$

Since S is based on Hu-moment norms and should therefore be minimised, the inverse of C was taken in those cases were histogram comparison returned a similarity function with opposite trend, i.e., for the Correlation and Intersection metrics. However, the Θ set was composed differently depending on the considered strategy. First, Θ_T included all θ_t, so that $\Theta_T = \{\theta_t : t = 1, \ldots, \sum_c \sum_i |V_i|\}$. Second, we averaged each θ by model (micro-average), creating Θ_Z and computing its arg min. For $z = 1, \ldots, N \sum_c |M_c|$:

$$\theta_z = \frac{\sum_{v_j \in V_i} \theta}{|V_i|} \tag{3}$$

Finally, each θ was averaged by class (macro-average), before being added to a Θ_C:

$$\theta_c = \frac{\sum_i \sum_{v_j \in V_i} \theta}{\sum_i |V_i|} \tag{4}$$

We assessed results for equal importance of contributing scores (i.e., $\alpha = 1$, $\beta = 1$), and then, increasing the relative importance of histogram comparison (i.e., $\alpha = 0.3$, $\beta = 0.7$), based on the prior batch of tests.

Cross-class cumulative accuracies are outlined in Table 2. Further class-wise details are left to Appendix A. In the hybrid trials, all combinations of shape-only and colour-only methods were evaluated, but we report only the configuration leading to the most consistent cumulative accuracy across all trials here, for the sake of brevity. The segmented objects in the NYUset were first matched against ShapeNetSet1 and, then, the latter was matched against different views collected under ShapeNetSet2, to control for the inherent characteristics of the NYU sample. In all described experiments, we took randomised label assignment as reference baseline.

3.3 Matching Feature Descriptors

Based on the results obtained in Section 3.2, we then tested whether relying on more general descriptions of image features would increase the accuracy of ShapeNet-based matching. To more easily assess the marginal variation introduced by these methods with respect to the prior trials, we directly compared ShapeNetSet2 against the reference ShapeNetSet1, in a more controlled scenario.

For all these trials, we relied on OpenCV built-in methods and used brute-force matching. Using FLANN-based matching for optimised nearest neighbour search did not lead to any performance gains, compared to the brute-force approach, most likely due to the fairly limited size of the input datasets. Therefore, we refrain from reporting results obtained with FLANN-based matching.

SIFT Firstly introduced by Lowe [16], the SIFT algorithm is based on the main rationale of describing images through scale-invariant keypoints. We used L2 norm as distance measure for the matching and trimmed the resulting matching keypoints to the second-nearest neighbour. A ratio test was then applied to select the best match among all reference 2D views at each iteration, as proposed in the original paper [16], setting the threshold to 0.75 and 0.5. respectively.

SURF was originally conceived for providing a more scalable alternative to SIFT, performing convolutions through square-shaped filters and, therefore, speeding up the computation [3]. Further, in SURF the keypoints are identified through maximising the determinant of the Hessian matrix for blob detection. We kept all the settings used for SURF in these trials and set the Hessian filter threshold to 400, to not overly reduce the output of the feature descriptor and thus retain sufficient richness in the representation.

ORB is another alternative approach to feature description implemented in the OpenCV Labs and proposed in [28]. ORB combines FAST for corner-based keypoint detection [27] with improved feature descriptors derived from BRIEF [7], to accommodate for rotation invariance. Since in BRIEF descriptors are parsed to binary strings to reduce their dimensionality, we used the Hamming distance instead of the L2 norm for this latter set of experiments.

During the evaluation, results were compared against randomised label assignment, similarly to 3.2. The obtained cumulative accuracy values are summarised in Table 3. The details on class-wise results obtained for the illustrated pipeline, are left to Appendix A.

Table 3: Cumulative (cross-class) accuracies after matching feature descriptors of views in ShapeNetSet1 (SNS1) against ShapeNetSet2 (SNS2).

Approach	Accuracy
Baseline	0.10
SIFT	0.25
SURF	0.22
ORB	0.25

3.4 Deep Neural Inexact Matching

Adapting the inexact-matching architecture proposed in [31] to our purpose, we implemented a Keras pipeline on top of a Tensorflow backend [9] for matching pairs of input images and binary classify them as similar or dissimilar. In [31] this method was designed to re-identify the same person across different frames; here we test for whether a similar approach can successfully scale to recognise diverse objects, in a task-agnostic fashion. This CNN-based architecture combines successive convolutions and pooling layers to both input images, sharing weights across the two input pipelines, drawing from the same rationale as Siamese Networks [6, 34]. Further, regions of pixels across the two image representations are compared so that a larger region is carried over from one image to another during the matching, hence explaining its *inexact* nature, as opposed to more traditional exact (Cosine-similarity) based matching techniques, as the one introduced in [6]. This strategy is expected to be more robust to wide viewpoint and illumination condition variations. Another property of normalized cross correlation matching is symmetry, making results independent from the ordering of images withing each couple the architecture is presented with and, thus, reducing the number of parameters needed in the subsequent layers. In addition to classic Siamese Networks, after similarity is computed as Normalized-X-Correlation, the output is further manipulated. Normalized-X-Corr tensors are in fact fed to two successive convolutional layers followed by Maxpooling, for dimensionality reduction and to summarize information gained on the neighbourhood of each pixel into a more dense representation [31]. Tensors are then fed to a fully-connected layer preceeding the final softmax layer to generate probabilities for the "similar" and "dissimilar" classes. In our Keras-based implementation, to achieve the desired dimensionality to feed to the softmax layer, we applied a dense layer and a flattening operation on cascade.

Further, the input RGB images were resized to 60x160x3 and the described model compiled using categorical crossentropy as loss function and Adam optimiser.

We used ShapeNetSet2 as baseline to form a training set, comprising of 9,450 RGB image pairs, with 52% being examples of similar images and the remainder 48% being labelled as dissimilar pairs. At training time, the learning rate was initialised to 0.0001 and its decay set to $1e-7$. Training samples were fed in batches of size 16 to run over up to 100 epochs. An early stopping condition was defined so that training would stop if the ϵ of loss decrease was lower than $1e-6$ for more than 10 subsequent epochs. As a result, training completed after 41 epochs, running on a NVIDIA Tesla P100 GPU.

Two different image sets were utilised on test: (i) 3,321 derived from image pairs in ShapeNetSet1, and (ii) 8,200 paired examples, obtained after matching 100 images from the NYUset (where 10 where randomly-picked from each of the 10 classes) with all views in ShapeNetSet1. The first experiment was conceived for checking on whether the Neural Network had learned to discriminate similar ShapeNet models, whereas the second one was meant to provide better insights on the results obtained in 3.2 and 3.3. Experimental results on both configurations, summed up in Table 4, are discussed in the following Section.

4 RESULTS AND DISCUSSION

Table 2 outlines how, with respect to cross-class, cumulative accuracy, all configurations outperformed random label assignment.
Interestingly enough, the weighted sum of shape and colour based scores were equal to the first-best results obtained with RGB histogram comparison alone. This could be due to a need to fine tune the α and β parameters, or it could also indicate that colour-based features are more prominent, when concluding about the recognised object. The latter hypothesis would align with another observation: shape-only trials led to the lowest cumulative accuracy values, among all tested setups. These observations hold also when controlling for the input data and comparing SNS2 against SNS1 instead of NYUset.
When taking a more careful look into class-wise results (as shown in Appendix A), one can notice how different approaches favoured different subsets of classes, when keeping the input data and boundary conditions constant, with only partial overlap across different pipelines and without any method completely outperforming the others in terms of cross-class consistency and robustness. On average, chairs were more highly recognised than other classes and most configurations led to unbalanced recognition, favouring certain classes at the expense of others. This applies also for the best case scenario, when matching SNS2 against SNS1, indicating that the inadequacy of the explored methods in robustly identifying all classes is not to be ascribed solely to the quality and characteristics of segmented areas within the NYU set. For instance, by looking at Table 8, it can be noted how the Paper and Window classes are not recognised in most of cases, even though, overall, the obtained performance was higher than in Table 7, due to the fact that all compared models belonged to ShapeNet.
Based on these initial results, when representing the input image in terms of SURF, SIFT or ORB based feature descriptors, we started from matching SNS2 against SNS1, to evaluate whether further application to the NYUset was worthwhile. As shown in Table 9, results obtained for the latter configurations were not sufficient and lower than the ones obtained with the hybrid strategies (Table 8), leading to cumulative accuracies in the range of 22% and 25% (Table 3).

Besides the feature engineering trials, we re-framed similarity learning also with respect to the Siamese-like architecture introduced in [31]. Our Keras implementation of the Normalized-X-Corr model was trained to learn an optimised representation from comparing pairs of images from SNS2, i.e., 9,450 pairs quite equally balanced between positive and negative examples, as introduced in Section 3.4. To achieve a more abstract and compact representation, where contributing features are not as clearly designed and discriminated as in the first set of experiments, the classification task was framed in binary terms at this stage. However, the tested setups led to unsatisfactory results that clearly indicate overfitting of the model (see Table 4). The even lower results obtained on the SNS1-derived test set can be further explained by looking at the unbalance between positive and negative examples, leading to a larger impact of false positives on the overall performance. The incidence of false positives is also partially caused by how the Normalized-X-Corr architecture was originally conceived [31], i.e., to match wider areas accommodating for varying viewpoint and luminance condition. However, the results obtained suggest that the chosen training set, feeding all possible permutations of couples in SNS2 to also minimise the number of required input labels, was not introducing sufficient variability, resulting in representations that did not generalise, even on unseen ShapeNet models. Further, the original use case of the exploited architecture was person re-identification, hinting towards further tweaking of the framework and hyperparamenter tuning to scale to multiple - and more diverse - object classes than simply human silhouettes.

Table 4: Class-wise Evaluation of our Keras implementation of Normalized-X-Corr, on the two labeled test sets.

Dataset	Measure	Similar	Dissimilar
ShapeNetSet1 pairs	Precision	0.09	0.00
	Recall	1.00	0.00
	F1-score	0.16	0.00
	Support	295	3026
NYU+ShapeNetSet1 pairs	Precision	0.51	0.00
	Recall	1.00	0.00
	F1-score	0.67	0.00
	Support	4160	4040

5 CONCLUSION

In this work, we tested for the relative importance of shape and colour based features in light of both cross-class and class-wise evaluation, and in experimental settings where we could control for the introduction of segmentation faults that would normally propagate from pre-processing steps. For this reason, we relied on a combination of a pre-segmented data set, i.e., RGB images from the large NYUDepth V2 set [21], and based similarity matching on a subset 2D models derived from the ShapeNet dataset [8]. Although features derived from comparing RGB histograms of the input images led, on average, to more consistent performances, none of the experimented pipelines ensured satisfactory results, in terms of robustness to class variation. Further, when adopting more general, scale, shift and rotation-invariant image representations, accuracy of classification by similarity matching was not sufficient, even when evaluating against alternative models all belonging to ShapeNet, hence controlling for other boundary conditions. Finally, the application of the Normalized-X-Architecture for inexact matching [31], formerly introduced in the context of person re-identification, led to overfitting and did not allow for subsequent application on unseen data sets and real-life settings.
All these findings confirmed the need for more scalable methods, capable of leveraging labelled and unlabelled data points, when learning object similarities with respect to diverse categories and taxonomies that imply high within-class heterogeneity. We therefore intend to modify the tested architecture accordingly, to improve its flexibility, while also increasing the heterogeneity of our datasets

(e.g., by representing a higher number of classes, and by augmenting the cardinality of each class), for further application on RGB frames captured by a mobile robot in a real-life scenario.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Center for Health Organization Transformation (CHOT) for graciously providing GPU computing resources exploited for some of the experiments presented in this work, and Rajeev Bhatt Ambati (Pennsylvania State University) for his patient and kind assistance on server-access-related matters.

REFERENCES

[1] Markus Bajones, David Fischinger, Astrid Weiss, Daniel Wolf, Markus Vincze, Paloma de la Puente, Tobias Körtner, Markus Weninger, Konstantinos Papoutsakis, Damien Michel, et al. 2018. Hobbit: Providing Fall Detection and Prevention for the Elderly in the Real World. *Journal of Robotics* 2018 (2018).

[2] Emanuele Bastianelli, Gianluca Bardaro, Ilaria Tiddi, and Enrico Motta. 2018. Meet HanS, the Health&Safety Autonomous Inspector. CEUR Workshop Proceedings.

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *European conference on computer vision*. Springer, 404–417.

[4] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. 2016. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*. Springer, 850–865.

[5] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. 7.

[6] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*. 737–744.

[7] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. 2010. Brief: Binary robust independent elementary features. In *European conference on computer vision*. Springer, 778–792.

[8] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

[9] François Chollet et al. 2015. Keras: Deep learning library for theano and tensorflow. *URL: https://keras. io/k* 7, 8 (2015).

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 248–255.

[11] Chester V Dolph, Loc Tran, and Bonnie D Allen. 2018. Towards Explainability of UAV-Based Convolutional Neural Networks for Object Classification. In *2018 Aviation Technology, Integration, and Operations Conference*. 4011.

[12] Thomas Fäulhammer, Rares Ambrus, Christopher Burbridge, Micheal Zillich, John Folkesson, Nick Hawes, Patric Jensfelt, and Marcus Vincze. 2017. Autonomous learning of object models on a mobile robot. *IEEE Robotics and Automation Letters* 2, 1 (2017), 26–33.

[13] Gabriele Ferri, Alessandro Manzi, Pericle Salvini, Barbara Mazzolai, Cecilia Laschi, and Paolo Dario. 2011. DustCart, an autonomous robot for door-to-door garbage collection: From DustBot project to the experimentation in the small town of Peccioli. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 655–660.

[14] Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*. Springer, 84–92.

[15] Ming-Kuei Hu. 1962. Visual pattern recognition by moment invariants. *IRE transactions on information theory* 8, 2 (1962), 179–187.

[16] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

[17] Nawel Medjkoune, Frédéric Armetta, Mathieu Lefort, and Stefan Duffner. 2017. Autonomous object recognition in videos using Siamese Neural Networks. In *EU-Cognition Meeting (European Society for Cognitive Systems) on" Learning: Beyond Deep Neural Networks"*.

[18] Martino Mensio, Emanuele Bastianelli, Ilaria Tiddi, and Giuseppe Rizzo. 2018. A Multi-layer LSTM-based Approach for Robot Command Interaction Modeling. *arXiv preprint arXiv:1811.05242* (2018).

[19] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[20] Christophe Mollaret, Alhayat Ali Mekonnen, Julien Pinquier, Frédéric Lerasle, and Isabelle Ferrané. 2016. A multi-modal perception based architecture for a non-intrusive domestic assistant robot. In *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on*. IEEE, 481–482.

[21] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*.

[22] Andreas Nüchter and Joachim Hertzberg. 2008. Towards semantic maps for mobile robots. *Robotics and Autonomous Systems* 56, 11 (2008), 915–926.

[23] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. *arXiv preprint* (2017).

[24] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 6 (2017), 1137–1149.

[26] Petar Ristoski and Heiko Paulheim. 2016. Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Web semantics: science, services and agents on the World Wide Web* 36 (2016), 1–22.

[27] Edward Rosten and Tom Drummond. 2006. Machine learning for high-speed corner detection. In *European conference on computer vision*. Springer, 430–443.

[28] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. In *Computer Vision (ICCV), 2011 IEEE international conference on*. IEEE, 2564–2571.

[29] Jian Shi, Yue Dong, Hao Su, and X Yu Stella. 2017. Learning non-lambertian object intrinsics across shapenet categories. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 5844–5853.

[30] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge.. In *AAAI*. 4444–4451.

[31] Arulkumar Subramaniam, Moitreya Chatterjee, and Anurag Mittal. 2016. Deep neural networks with inexact matching for person re-identification. In *Advances in Neural Information Processing Systems*. 2667–2675.

[32] Ilaria Tiddi, Emanuele Bastianelli, Gianluca Bardaro, Mathieu d'Aquin, and Enrico Motta. 2017. An ontology-based approach to improve the accessibility of ROS-based robotic systems. In *Proceedings of the Knowledge Capture Conference*. ACM, 13.

[33] Yu Xiang and Dieter Fox. 2017. DA-RNN: Semantic mapping with data associated recurrent neural networks. *arXiv preprint arXiv:1703.03098* (2017).

[34] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Deep metric learning for person re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 34–39.

[35] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois R Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, et al. 2018. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1–8.

A TABLES OF RESULTS

The results enclosed in Tables 5, 6, and 7 refer to all exploratory tests run when images from the NYUset were matched against ShapeNetSet1 (SNS1) and evaluated class by class. Table 8 summarizes results obtained when combining shape-only and color-only scores computed on images from ShapeNetSet2 (SNS2) and performing the matching against instances of ShapeNetSet1 (SNS1). Similarly, class-wise results in Table 9 refer to matching feature descriptors of views in SNS1 against descriptors of models in SNS2.

Table 5: Class-wise results obtained when matching only based on shape.

Approach	Measure	Chair	Bottle	Paper	Book	Table	Box	Window	Door	Sofa	Lamp
Baseline	Accuracy	0.15600	0.10543	0.11899	0.10132	0.11846	0.08948	0.08104	0.07241	0.09899	0.09414
	Precision	0.02250	0.01399	0.01356	0.01110	0.01240	0.00822	0.00721	0.00534	0.00707	0.00649
	Recall	0.15600	0.10543	0.11899	0.10132	0.11846	0.08948	0.08104	0.07241	0.09899	0.09414
	F1-score	0.03932	0.02470	0.02434	0.02002	0.02245	0.01506	0.01324	0.00994	0.01319	0.01214
L1	Accuracy	0.25900	0.39565	0.04810	0.00132	0.15702	0.00471	0.00000	0.00783	0.36768	0.06276
	Precision	0.03735	0.05249	0.00548	0.00014	0.01644	0.00043	0.00000	0.00058	0.02625	0.00433
	Recall	0.25900	0.39565	0.04810	0.00132	0.15702	0.00471	0.00000	0.00783	0.36768	0.06276
	F1	0.06529	0.09269	0.00984	0.00026	0.02977	0.00079	0.00000	0.00107	0.04900	0.00809
L2	Accuracy	0.08500	0.81413	0.00759	0.00132	0.03581	0.00157	0.00000	0.00978	0.24444	0.02929
	Precision	0.01226	0.10802	0.00087	0.00014	0.00375	0.00014	0.00000	0.00072	0.01745	0.00202
	Recall	0.08500	0.81413	0.00759	0.00132	0.03581	0.00157	0.00000	0.00978	0.24444	0.02929
	F1	0.02143	0.19073	0.00155	0.00026	0.00679	0.00026	0.00000	0.00134	0.03258	0.00378
L3	Accuracy	0.32700	0.46413	0.04557	0.00395	0.07989	0.01099	0.00162	0.00978	0.32121	0.15690
	Precision	0.04716	0.06158	0.00519	0.00043	0.00836	0.00101	0.00014	0.00072	0.02293	0.01082
	Recall	0.32700	0.46413	0.04557	0.00395	0.07989	0.01099	0.00162	0.00978	0.32121	0.15690
	F1	0.08243	0.10873	0.00932	0.00078	0.01514	0.00185	0.00026	0.00134	0.04281	0.02024

Table 6: Class-wise results obtained when comparing RGB histograms (baseline same as Table 5)

Matching metric	Measure	Chair	Bottle	Paper	Book	Table	Box	Window	Door	Sofa	Lamp
Correlation	Accuracy	0.56500	0.04130	0.20506	0.09211	0.03581	0.06750	0.08104	0.03327	0.14949	0.12971
	Precision	0.08148	0.00548	0.02336	0.01010	0.00375	0.00620	0.00721	0.00245	0.01067	0.00894
	Recall	0.56500	0.04130	0.20506	0.09211	0.03581	0.06750	0.08104	0.03327	0.14949	0.12971
	F1	0.14243	0.00968	0.04195	0.01820	0.00679	0.01136	0.01324	0.00457	0.01992	0.01673
Chi-square	Accuracy	0.48900	0.00000	0.00000	0.00921	0.13085	0.04710	0.44408	0.00196	0.00000	0.23431
	Precision	0.07052	0.00000	0.00000	0.00101	0.01370	0.00433	0.03952	0.00014	0.00000	0.01615
	Recall	0.48900	0.00000	0.00000	0.00921	0.13085	0.04710	0.44408	0.00196	0.00000	0.23431
	F1	0.12327	0.00000	0.00000	0.00182	0.02480	0.00792	0.07257	0.00027	0.00000	0.03022
Intersection	Accuracy	0.57200	0.19565	0.30886	0.01447	0.03581	0.01884	0.01945	0.04892	0.38182	0.06485
	Precision	0.08249	0.02596	0.03519	0.00159	0.00375	0.00173	0.00173	0.00361	0.02726	0.00447
	Recall	0.57200	0.19565	0.30886	0.01447	0.03581	0.01884	0.01945	0.04892	0.38182	0.06485
	F1	0.14419	0.04584	0.06318	0.00286	0.00679	0.00317	0.00318	0.00672	0.05088	0.00836
Hellinger	Accuracy	0.53800	0.08370	0.38228	0.01974	0.03168	0.03925	0.44895	0.05284	0.24242	0.05649
	Precision	0.07759	0.01110	0.01110	0.00216	0.00332	0.00361	0.03995	0.00389	0.01731	0.00389
	Recall	0.53800	0.08370	0.38228	0.01974	0.03168	0.03925	0.44895	0.05284	0.24242	0.05649
	F1	0.13562	0.01961	0.02158	0.00390	0.00601	0.00660	0.07337	0.00725	0.03231	0.00729

Table 7: Class-wise results obtained when combining L3 norm-based Hu moment matching with Hellinger distance-based RGB histogram comparison, when class labels are determined based on minimizing: (i) the weighted sum of scores (ii) the micro-average of scores and (iii) the macro-average of scores. We report the weight configuration that ensured the most consisted results, among the tested ones, i.e., setting $\alpha = 0.3$, $\beta = 0.7$. See Table 5 for reference baseline.

Argmin function	**Measure**	Chair	Bottle	Paper	Book	Table	Box	Window	Door	Sofa	Lamp
Weighted Sum	Accuracy	0.65300	0.14891	0.12658	0.00526	0.10055	0.02512	0.29011	0.05871	0.28081	0.20921
	Precision	0.09417	0.01976	0.01442	0.00058	0.01053	0.00231	0.02581	0.00433	0.02005	0.01442
	Recall	0.65300	0.14891	0.12658	0.00526	0.10055	0.02512	0.29011	0.05871	0.28081	0.20921
	F1	0.16461	0.03489	0.02589	0.00104	0.01906	0.00423	0.04741	0.00806	0.03742	0.02698
Micro-average	Accuracy	0.37800	0.13587	0.18861	0.02105	0.04821	0.07064	0.37925	0.10568	0.22626	0.07741
	Precision	0.05451	0.01803	0.02149	0.00231	0.00505	0.00649	0.03375	0.00779	0.01615	0.00534
	Recall	0.37800	0.13587	0.18861	0.02105	0.04821	0.07064	0.37925	0.10568	0.22626	0.07741
	F1	0.09529	0.03183	0.03858	0.00416	0.00914	0.01189	0.06198	0.01451	0.03015	0.00998
Macro-average	Accuracy	0.39000	0.15543	0.39241	0.00000	0.11846	0.06750	0.00000	0.00000	0.29495	0.05649
	Precision	0.05624	0.02062	0.04471	0.00000	0.01240	0.00620	0.00000	0.00000	0.02106	0.00389
	Recall	0.39000	0.15543	0.39241	0.00000	0.11846	0.06750	0.00000	0.00000	0.29495	0.05649
	F1	0.09831	0.03641	0.08027	0.00000	0.02245	0.01136	0.00000	0.00000	0.03931	0.00729

Table 8: Similarly to Table 7, but matching SNS2 against SNS1.

Argmin function	**Measure**	Chair	Bottle	Paper	Book	Table	Box	Window	Door	Sofa	Lamp
Weighted Sum	Accuracy	0.90	0.10	0.00	0.20	0.30	0.10	0.00	0.50	0.40	0.70
	Precision	0.09	0.01	0.00	0.02	0.03	0.01	0.00	0.05	0.04	0.07
	Recall	0.90	0.10	0.00	0.20	0.30	0.10	0.00	0.50	0.40	0.70
	F1	0.16	0.02	0.00	0.04	0.05	0.02	0.00	0.09	0.07	0.13
Micro-average	Accuracy	0.80	0.10	0.00	0.30	0.20	0.20	0.10	0.60	0.30	0.20
	Precision	0.08	0.01	0.00	0.03	0.02	0.02	0.01	0.06	0.03	0.02
	Recall	0.80	0.10	0.00	0.30	0.20	0.20	0.10	0.60	0.30	0.20
	F1	0.15	0.02	0.00	0.05	0.04	0.04	0.02	0.11	0.05	0.04
Macro-average	Accuracy	0.70	0.60	0.00	0.00	0.10	0.10	0.00	0.00	0.60	0.10
	Precision	0.07	0.06	0.00	0.00	0.01	0.01	0.00	0.00	0.06	0.01
	Recall	0.70	0.60	0.00	0.00	0.10	0.10	0.00	0.00	0.60	0.10
	F1	0.13	0.11	0.00	0.00	0.02	0.02	0.00	0.00	0.11	0.02

Table 9: Class-wise results obtained when matching feature descriptors derived from SIFT, SURF and ORB. We report the configurations that ensured the most consistent results, among the tested ones, i.e., for a ratio test threshold set to 0.5.

Approach	**Measure**	Chair	Bottle	Paper	Book	Table	Box	Window	Door	Sofa	Lamp
SIFT	Accuracy	0.30	0.30	0.00	0.40	0.00	0.40	0.30	0.20	0.30	0.30
	Precision	0.03	0.03	0.00	0.04	0.00	0.04	0.03	0.02	0.03	0.03
	Recall	0.30	0.30	0.00	0.40	0.00	0.40	0.30	0.20	0.30	0.30
	F1	0.05	0.05	0.00	0.07	0.00	0.07	0.05	0.04	0.05	0.05
SURF	Accuracy	0.70	0.10	0.00	0.10	0.10	0.00	0.30	0.30	0.30	0.30
	Precision	0.07	0.01	0.00	0.01	0.01	0.00	0.03	0.03	0.03	0.03
	Recall	0.70	0.10	0.00	0.10	0.10	0.00	0.30	0.30	0.30	0.30
	F1	0.13	0.02	0.00	0.02	0.02	0.00	0.05	0.05	0.05	0.05
ORB	Accuracy	0.10	0.70	0.00	0.20	0.10	0.00	0.30	0.20	0.40	0.50
	Precision	0.01	0.07	0.00	0.02	0.01	0.00	0.03	0.02	0.04	0.05
	Recall	0.10	0.70	0.00	0.20	0.10	0.00	0.30	0.20	0.40	0.50
	F1	0.02	0.13	0.00	0.04	0.02	0.00	0.05	0.04	0.07	0.09