Non-Parametric Spatial Context Structure Learning for Autonomous Understanding of Human Environments

Akshaya Thippur† Johannes Stork† and Patric Jensfelt†

***Abstract*— Autonomous scene understanding by object classification today, crucially depends on the accuracy of appearance based robotic perception. However, this is prone to difficulties in object detection arising from unfavourable lighting conditions and vision unfriendly object properties. In our work, we propose a spatial context based system which infers object classes utilising solely structural information captured from the scenes to aid traditional perception system.**

Our system operates on novel spatial features (IFRC) that are robust to noisy object detections; It also caters to on-the-fly learned knowledge modification improving performance with practise. IFRC are aligned with human expression of 3D space, thereby facilitating easy HRI and hence simpler supervised learning. We tested our spatial context based system to successfully conclude that it can capture spatio structural information to do joint object classification to not only act as a vision aide, but sometimes even perform on par with appearance based robotic vision.

Keywords: structure learning, spatial relationships, lazy learners, autonomous scene understanding

I. Introduction

In the near future, autonomous helper robots will need to automatically learn about their specific environment, the human members and their activities over time. It is vital for these systems to process streams of noisy image data in day-to-day dynamic environments, be able to model, abstract and generalise from little data and maintain long-term knowledge with plasticity and transferability. The robot needs to understand 'common factors' in scenarios and adaptively apply what it has previously learnt.

The above functionalities fundamentally depend on reliable recognition of the plethora of objects in the environment. Appearance based robotic perception are troubled by, variety in class instances, bad segmentation due to bad lighting and real-world confusing clutter in environments. The challenges faced by state-of-the-art vision based perception systems for scene understanding [1], [2], [3], can be eased using spatio-temporal context information [4], [5], [6]. Consider Bob the robot in Fig.1; it guesses that the object is a keyboard because of its spatial context which is composed of the individual spatial relationship features (SRFs) of that object of interest with respect to other objects in the scene. We believe that long-term learning can be more effective by extracting and compiling structures in environments. To this end, we propose a novel spatial context structure learning system to improve robotic perception.

†RPL (CVAP), KTH Royal Institute of Technology, Sweden: akshaya, jastork, patric@kth.se

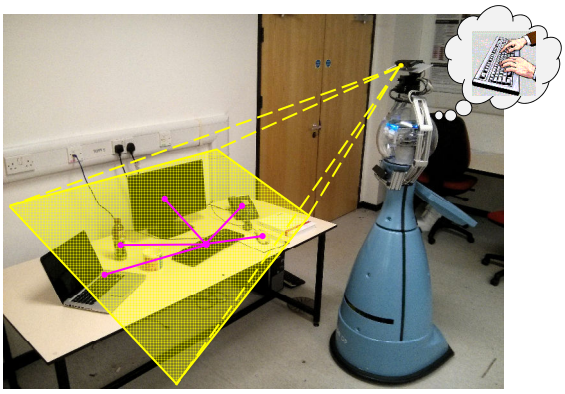


Fig. 1: 'Bob' seen here, is a SCITOS-G5 mobile robot platform which is observing a desktop (●) using a 3D camera mounted atop his head. He recognises a keyboard based on its spatial context – comprised of inter-object spatial relationships (➳).

Perception systems using appearance and spatial context have been introduced in the work of [7], [15]; The authors in [7] compare the efficacies of two different spatial context based systems which aid appearance based systems on the task of *Joint Object Classifcation* in desktop settings, where the challenge is to assign correct labels to all detected objects in the scene. The performance of their perception improves because of inculcating spatial context information into the reasoning. However, the features they extract depend critically on a spatial origin and their learners have drawbacks with respect to knowledge amendment and speed of reasoning.

In this paper we propose a knowledge modification friendly, faster and significantly better *Context Comprehension System* (CCS) to aid robotic perception. This utilises only inter-object spatial relationships to independently perform joint object classification. Our system only requires rough pose and size detections of objects in a 3D observation of the scene in question. We introduce a novel set of SRFs, called *Intrinsic Frame of Reference Calculi* (IFRC), which is computed between all object pairs. Subsequently, our CCS, learns from ground truth data and employs a weighted k-Nearest Neighbour (kNN) based technique to learn features→class-label mappings. Voting schemes along with a multiclass kNN classifier infer the class labels of unknown objects in a novel scene.

Through our experiments we show that a *joint* second-order treatment of the labelling problem can successfully aid

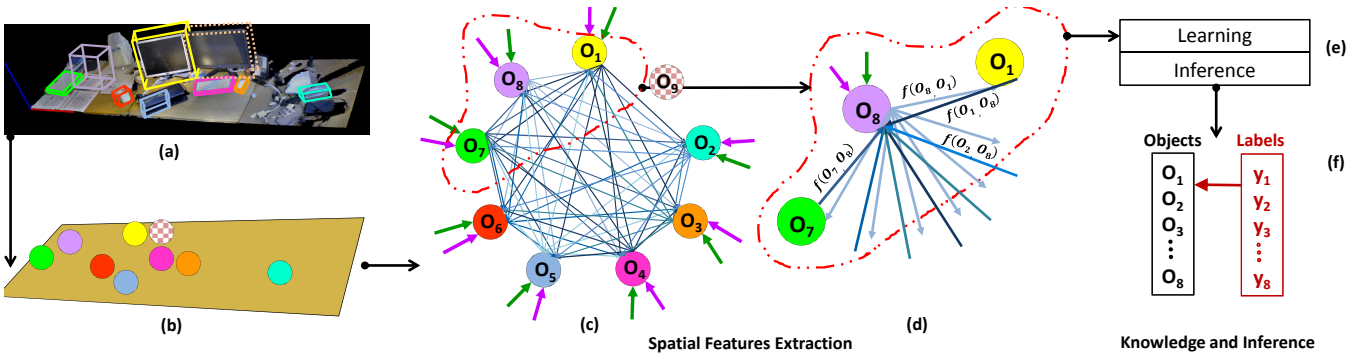


Fig. 2: **Object recognition schematic of context comprehension system** (L to R). (a) In a 3D image of a desktop scene, objects are detected as oriented minimum bounding boxes (different colors). (b) Each object is a node which has no label but only measurements of position, pose and rough size in the global frame of the desk. (c) In the Spatial Features Extraction part, IFRC are computed from each object to every other object to capture scene structure (in shades of blue). (d) takes a closer look at a component of the scene structure stemming from object O_8. $f(O_8, O_t) : \forall t \neq 8$ (⟶) are the *outgoing* IFRC vectors of all other objects with respect to landmark O_8; and $f(O_l, O_8) : \forall l \neq 8$ (different shades of blue) are the *incoming* IFRC vectors of trajector O_8 with respect to all other objects. (e) Such training data is provided to the Knowledge and Inference part which learns a knowledge bank. When such features are computed on a novel scene, a KNN based inference scheme is employed to estimate object class labels as in (f). In (c,d) the arrows ⟶ corresponding to appearance based cues and ⟶ to single object spatio-temporal statistics, schematically show how our CCS could integrate into a robotic perception system as a whole.

and in some cases even perform on par with state-of-the-art appearance based systems. A system using this treatment can provide object classifications when the appearance based system can provide only object segments due to systemic or environment problems. Subsequently, we also show that a simpler and faster *approximate* second-order treatment provides inferences which can be incorporated into robotic perception. Inference using this treatment can be deployed on robotic platforms to work in tandem with real-time appearance based perception, to improve wholistic performance.

Thus our contributions are: (1) A novel, robust, human robot interaction (HRI) friendly spatial relation set – IFRC. (2) A context comprehension system which offers accurate object labellings despite having a low complexity. (3) Experimental evidence to support the hypothesis –"there is sufficient information contained in underlying structures to use them independently for scene understanding". Note: Our context comprehension system is proposed to support appearance based perception systems, not to replace them.

II. Related Work

Traditionally, robotic vision systems perceive, understand and reason about their surroundings using appearance based features [8], [9]. However, contextual information has been used for improving object recognition [10], [11], [12], activity recognition [13], [14], [15] and scene understanding [16], [17]. Since the components of human environments – backgrounds, objects, humans and other living beings – rarely occur in isolation, it is possible to obtain better recognition, by appending evidence from their spatio temporal contexts to the appearance based perception systems.

Spatio temporal context information can be a simplistic encoding of neighbourhood measurements of the already used appearance based features [18], [19] or statistics of context co-occurrences [20], [21]. On the other hand, contextual information can be captured using sophisticated approaches involving dedicated feature sets utilised by specialised inference methods. Spatial relations can be encoded using pairwise geometric calculi [22], [23] and temporal context can be characterised using cues from time series analysis [24]. Such contextual information can also be learned [25], [26], [27], [28] using the scope of the research problem rather than imposing pre-defined relations.

The work in [7] shows that robotic perception systems depending on coarse SRFs are helpful for early and mid stage learning in long-term autonomy when there is low availability of training data. The work in [29] presents a generic superset of spatial relationship concepts using combinations of which, any spatial relationship between objects in any environment can be described. In our work, we learn underlying structures using spatial relations and characterise the environment and its components with such structures.

Scene understanding achieved by abstracting over object recognitions has improved with the use of SRFs. In [30], Gupta et.al. use 3D geometry constraints and understand scenes represented in real-world images in terms of adjacent semantic 3D blocks. Choi et.al. in [31], use camera geometry to obtain best fit bounding boxes of component objects in a 2D scene. These box-objects are then clustered and learned in the form of semantic trees to provide spatial relation based structural information for supporting inference. Lin et.al. in [32] use object detections and compute geometric potential functions between pairs which reflect spatial relationships. These potentials are factored into a conditional random field for object inference. In our work we propose the use of IFRC, a spatial relations feature set, which is vitally grounded in human linguistics to facilitate HRI and yet preserve the properties required to capture structures to achieve joint object recognition and provide for environment understanding; Work in [33], [34] explore the uses of inculcating human language into environment descriptions for robots.

Our CCS uses a non-parametric approach for structure learning and multiclass inference [35] for joint object recognition. It employs a lazy learner and populates a metric-subspace with labelled exemplars. The work in [36] ap-

plies kernel density estimation to separate background and foreground in an image. We apply similar principles in the learned metric-subspace to focus on the queried neighbourhood and obtain a radial basis function weighted nearest neighbour estimate. Another approach would be to do a fixed number nearest neighbour query like the work in [37], [38]. Chiang et.al. [39] use a ranking of the queried neighbours for multiclass classification. We employ a weighted KNN based approach to query the metric-subspace for multiclass classification and separately to provide label-belief factors for maximum a-posteriori reasoning.

The authors in [40], [7] propose similar approaches which operate on SRF sets for object recognition and scene understanding. These feature measurements crucially depend on the location of the observer and the size of objects, which are usually measured with uncertainty. CCS operates on IFRC which are all relative measurements. They only need rough object sizes and a general direction of an observer for an initialisation. The inference systems used in [7], [41], [42] have probabilistic modelling, which involve learning parametric distributions for class labels over extracted features. Modelling this ensemble of distributions requires large amounts of ground truth data. In contrast, CCS uses non-parametric lazy learners which do not need intensive training.

Learned knowledge might need modification in a long-term setting. For probabilistic methods in [7], [41], [42], this can be achieved only by expensive re-training. In contrast, the CCS' lazy learner disadvantages help on-the-fly knowledge modification by having virtually no training costs.

III. Spatial Features Extraction

We introduce a set of spatial relations called the Intrinsic Frame of Reference Calculi. IFRC are geometric definitions grounded in human linguistic concepts which are commonly used to describe 3D space across many vernaculars [44]. IFRC is a collection of spatial geometry predicate measurements between a pair of objects e.g. How much right-of the monitor is the mug? IFRC is comprised of 4 directional predicates, 2 proximity predicates and 1 overlap predicate.

We formulated these definitions after we conducted a perception test on 12 subjects in 3 languages. They described controlled real world object settings on desktops using relational prepositions of their choice. The object configurations in the tests were adaptively modified based on subjects' previous answers to approximately trace out the IFRC boundaries. As IFRC are grounded in linguistic concepts, it becomes a WYSIWYG language, facilitating easier human robot interaction (HRI). A long-term autonomous robot could thus be trained by a human supervisor without robotics related expertise.

IFRC are robust to noise in size and shapes of object detections and are independent of object appearances in contrast to traditional perception systems (see§VI). In comparison, appearance based perception operate on keypoints, colors, textures and salient object parts which are dependent on segmentation fidelity and stable lighting. Many a time, due to faulty segmentation, chunks of the object go missing – which makes measurements of size also unreliable. IFRC will describe the same objects devoid of such critical dependencies. Lastly, IFRC are quickly calculable using closed form geometric equations.

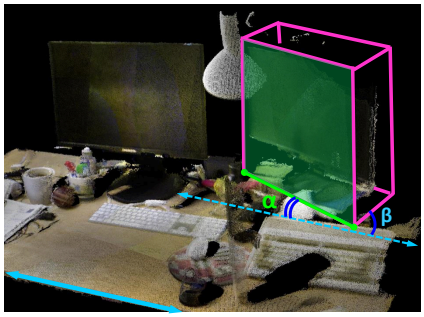


Fig. 3: The *Principal Face* (■) and the *Principal Edge* as its lower edge (•–•) are determined for this monitor using angles $\alpha < \beta$ subtended to front edge of desk (–). (§III)

We make the following assumptions and conventions for IFRC definitions:

1) Real-world scenes can be compelled to a global definition of origin and coordinate directions. For example: All objects stand on X-Y plane with their heights in the positive Z direction. The origin is always at the nearer left corner of the desktop scene.
2) An oriented minimum bounding box in 3D space can be computed for every object (Fig.2). From here onwards, when we use the term 'object', it refers to the oriented minimum bounding box of the object. This relates to a rough object detection, where there is unreliable capturing of shape, size and appearance based features.
3) A *Principal Face* of higher interest can be computed for every object (Fig.3). The principal face of objects are calculated as that plane which makes the least angle with the front of the desktop. The X-Y projected edge of the principal face, closest to the interacting human, is called the *Principal Edge*.
4) The linear measures of an object in the X-Y plane, parallel to the principal edge is *length*, perpendicular to it is *width* and along the Z-axis is *height*.
5) The volume (area, in X-Y) confined by the object's oriented minimum bounding box is defined as the *inside* of the object. The rest is *outside* of the object.
6) For the rest of the paper we deal with objects only using their X-Y projections, to account for incomplete, unreliable object segmentations.

The directional IFRC spatial relations of a *Trajector* (O_t) object wrt. a *Landmark* (O_l) object is defined from the 'perspective of the landmark' or its *intrinsic frame-of-reference*. Given a landmark and a trajector we can algorithmically extract the directional *IFRC feature vector*, $f(O_l, O_t) \in I^7$ where, $I = [0, 1]$ in $\mathbb{R}$.

***Directional Predicates**:* The space around the landmark is demarcated into *Front, Behind, Left* and *Right* fields (Fig.4) based on the perception tests' results. A space expanding outside of the object in the direction of a vector pointing toward the principal face from the centroid of the object is *Front Field*; And a space expanding similarly in a direction opposite to this vector is the *Behind Field*. These fields extend infinitely along the width of the object but along the

length, they are bound by one extra length of the object on either side of it as shown in Fig. 4. The *Left Field* and *Right Field* are defined to the landmark's left and right directions and have boundaries consequent to the definitions of front and behind fields.

The measure of front, behind, left and right of a trajector is the fraction of its total area (A^t_{total}) lying in those corresponding fields ($A^t_{\text{front field}}$) as shown in Fig.4. Thus,

$$\mathcal{F} \;\coloneqq\; A^t_{\text{front field}}/A^t_{\text{total}} \quad \ldots \text{similarly for } \mathcal{B}, \mathcal{L}, \mathcal{R} \quad \text{and } \mathcal{B}, \mathcal{F}, \mathcal{L}, \mathcal{R} \in I. \tag{1}$$

Proximity Predicates: In IFRC, we define two flavours of non-directional, proximity predicates or *Nearness* as, functions of Euclidean distances between objects, differing in the rough size contexts they use. For both, we first find the direction of the shortest distance between landmark and trajector (ρ_P).

- *Nearness-Projected* ($\mathcal{N}_\mathcal{P}$): We project the objects along the direction of ρ_P and find the size context distance D_P as shown in Fig.5. We define,

$$\mathcal{N}_\mathcal{P} \coloneqq e^{-\rho_P/D_P} \quad \text{where, } \mathcal{N}_\mathcal{P} \in I. \tag{2}$$

  We interpret $\mathcal{N}_\mathcal{P}$ as proximity in the context of size of the objects when viewed in ρ_P-centric perspective.
- *Nearness-Diagonal* ($\mathcal{N}_\mathcal{D}$): We calculate another size context distance D_D by summing the diagonals of both objects: $D_D = d_l + d_t$. We define,

$$\mathcal{N}_\mathcal{D} \coloneqq e^{-\rho_P/D_D} \quad \text{where, } \mathcal{N}_\mathcal{D} \in I. \tag{3}$$

  $\mathcal{N}_\mathcal{D}$ in contrast to $\mathcal{N}_\mathcal{P}$ considers the proximity of objects in the contexts of their absolute sizes.

Note: both $\mathcal{N}_\mathcal{P}$ and $\mathcal{N}_\mathcal{D}$ would assume a maximum value of 1 if the objects are touching or overlapping with each other.

Overlap Predicate: We define the amount of overlap $\mathcal{O}$ as the fraction of the total area (A^t_{total}) of the trajector lying *inside* the landmark (A^t_{landmark}),

$$\mathcal{O} \coloneqq A^t_{\text{landmark}}/A^t_{\text{total}} \quad \text{where, } \mathcal{O} \in I \tag{4}$$

This predicate is directional. Consider the example of a mug (obj-A) placed on top of a laptop (obj-B). It could be that $\mathcal{O}_{B\to A} = 1$ but $\mathcal{O}_{A\to B} = 0.2$

Concatenating all of these predicates, we obtain the directional IFRC feature vector extending from the landmark to the trajector as:

$$f(O_l, O_t) \coloneqq [\mathcal{B}, \mathcal{F}, \mathcal{L}, \mathcal{R}, \mathcal{N}_\mathcal{P}, \mathcal{N}_\mathcal{D}, \mathcal{O}]^\top \text{ where, } f \in I^7 \tag{5}$$

IV. Context Comprehension System

We give an overview of our CCS and subsequently elaborate on its different parts in the following sections (Fig.2).

Our CCS begins with the Spatial Feature Extraction part. It extracts IFRC vectors between objects for every scene S as $F_S = \{f_{lt} : \forall l, t \in \{1 : N_S\}\}$ using the definitions (§III).

The Knowledge and Inference part uses lazy machine learning to learn spatial structural knowledge from the environment scenes (§IV-A). For every training scene, IFRC

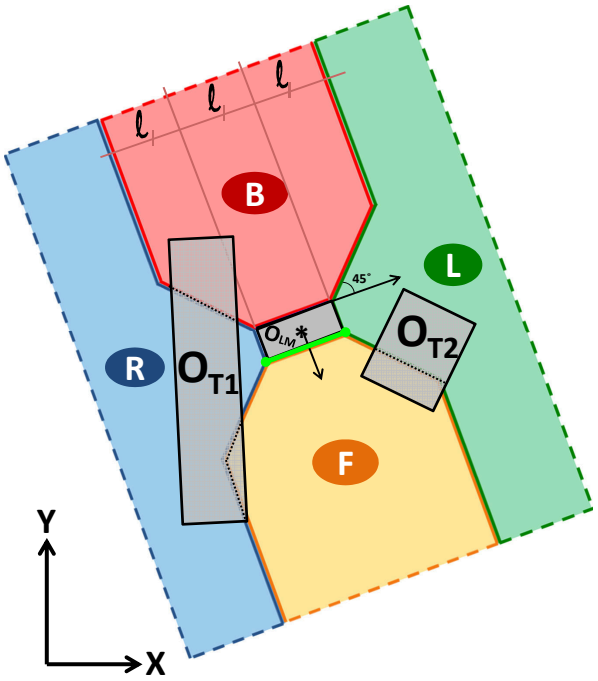


Fig. 4: IFRC *fields* wrt. landmark object O_{LM} with centroid (∗). IFRC vectors are computed to trajector objects O_{T1} and O_{T2}. Principal edge (●–●) of O_{LM} determines the fields: behind–B, front–F, left–L and right–R. Here, directional predicates [B,F,L,R] of IFRC vectors to T1 and T2 are $f_{LT_1} = [0.3, 0.1, 0.0, 0.6]$ and $f_{LT_2} = [0.0, 0.3, 0.6, 0.0]$. The length ℓ of O_{LM} is used to define the behind and forward fields as shown. (§III)

examples are extracted for each possible object pair and compiled into a *Knowledge Bank* (Ξ) with their corresponding object class labels.

The inference part assumes we have the following set of K object classes in our universe $\mathbb{C} = \{C_1, C_2, \cdots, C_K\}$; Given a scene S with N_S unknown objects, it finds the best class label sequence estimate: $Y^* = \{y^*_1, ..., y^*_{N_S}\}$ where $y^*_i \leftarrow C_i \in \mathbb{C}$. The inference queries and considers the nearest neighbours and their labels for maximum a-posteriori reasoning and voting schemes which provide the required class label estimates. CCS treats joint object classification in either of two ways: In the *Joint Second-Order* treatment (§IV-B), we reason about all the object labels by jointly focusing on all possible object pairs; In the *Approximate Second-Order* treatment (§IV-C), reasoning is relaxed to one object at a time, still strictly keeping with its collective pairwise spatial context.

A. Learning and Querying

The CCS learns non-parametric models using training data of objects in scenes. For every scene, IFRC are extracted for every possible directional object pair and stored along with the true class labels for those objects. This builds a knowledge bank (Ξ) as,

$$\Xi \leftarrow \{(f_{lt}, C_l, C_t) : \forall\, l, t \in \mathbb{O}\}_S \,\forall\, S \in \mathbb{S} \tag{6}$$

where C_l, C_t are true class labels for landmark and trajector objects respectively and $\mathbb{S}$ is the training set of scenes.

The Ξ is used differently by the two treatments. When our CCS encounters a novel scene, IFRC vectors are extracted

between all object pairs F_S and fed to the inference part. CCS queries Ξ for nearest neighbours $\forall f \in F_S$ and then uses either of the problem treatments to infer all object labels.

We implemented a kernel density estimate [36] based retrieval (RBNN) with a radial basis function weighting, parametrised by its radius r. We also implemented an Euclidean metric based KNN retrieval classifier which queries for k voting members [43]. Voting weights for these k members are computed using an exponentially decreasing function of Euclidean distance and inverse of class occurrence frequency to undo data bias. In RBNN query, the k number of nearest neighbour voting members varies depending on the density of data distribution at the queried point in the metric-subspace of Ξ. For simplicity of notation we will use k for number of queries for both types of queries.

B. Joint Second-Order Treatment

For a novel scene S, for each $f_{ij} \in F_S$ our CCS retrieves voting members and their corresponding class labels C_i, C_j from Ξ. The retrieved voting members are compiled into a normalised factor ϕ in accordance to their weights and parametrised by f_{ij}:

$$\phi_{f_{ij}}(y_i, y_j) \qquad \forall\, y_i, y_j \in \mathbb{C} \times \mathbb{C} \tag{7}$$

Such a factor, gives an empirical probability distribution over all possible label assignments to that object pair: $P(y_i{=}C_p, y_j{=}C_q|f_{ij})$ where $\forall p, q \in \{1:K\}$. A unimodal maximum value is expected for the target labelling of the objects (O_i, O_j) at $(y_i, y_j) = (C_p, C_q)_{\text{target}}$ in $\phi_{f_{ij}}$.

We do joint object classification of the N_S by considering all the IFRC vectors together to get a best labelling sequence. We suppose an empirical probability score $\tilde{P}$ for an estimated labelling sequence Y to be proportional to all $2\binom{N_S}{2}$ object-pairwise factors $\phi(y_i, y_j)$ parametrised by their corresponding computed IFRC f_{ij}

$$\tilde{P}(Y) \propto \prod_{i,j \in \{1:N_S\}, i \neq j} \phi_{f_{ij}}(y_i, y_j) \qquad \text{where,} \quad \phi_{f_{ij}}(y_i, y_j) : K \times K \rightarrow \mathbb{R}^{K^2} \tag{8}$$

These factors are then combined as in Eqn.9 to search for the label assignment sequence Y^* which maximises the empirical probability $\tilde{P}$.

$$\begin{aligned} Y^* &= \arg\max_Y \tilde{P}(Y) \qquad \text{or,} \\ Y^* &= \arg\max_Y \sum_{i,j \in \{1:N_S\}, i \neq j} \log(\phi_{f_{ij}}(y_i, y_j)) \end{aligned} \tag{9}$$

Our joint second-order treatment of joint object classification could be realised using a factor graph. However, to keep our results away from implementation biases, we solve the above optimisation problem using a pruned depth-first search algorithm.

C. Approximate Second-Order Treatment

Approximate joint second-order treatment is a relaxation of the joint second-order treatment, which has a lower time complexity at the expected cost of some accuracy. Here, N_S objects in a scene S are labelled by considering each object independent of other objects. We take an object $O_i \in \mathbb{O}_{1:N_S}$ in the scene and extract IFRC features with respect to all other objects O_j in the scene as the landmark in-turn: $\{f_{ji} \forall j = \{1:N_S\}, j \neq i\}$. These $\{f_{ji}\}$ are used to infer the object label y_i^* for O_i. This process is repeated for all objects in-turn to obtain the estimated labelling $Y^* = \{y_1^*, ..., y_{N_S}^*\}$. The features in use are still solely, pairwise IFRC vectors; It is still a second-order treatment because the mechanism to infer the final label sequence is piecewise independent (approximate) as opposed to the combined undertaking by the joint second-order treatment.

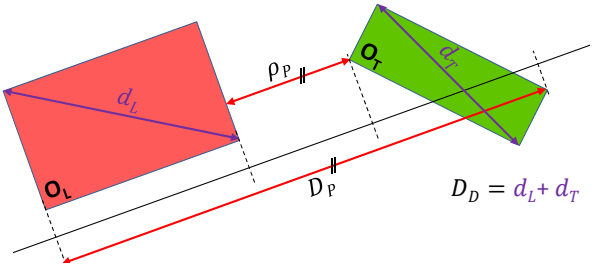


Fig. 5: Proximity components $\mathcal{N}_\mathcal{P}$ and $\mathcal{N}_\mathcal{D}$ are computed between Landmark O_L and Trajector O_T as using the rough contextual size measurements: D_P and D_D. (§III)

In this treatment, for an object O_i, CCS performs KNN or RBNN queries for k nearest neighbour voting members for each f_{ji} in Ξ (§IV-A). Only their weighed trajector labels C_t are considered for voting (Eqn.6). These votes are max-pooled to provide a class label estimate y_{ji} for O_i from O_j:

$$y_{ji}^* = \arg\max_C \ \text{HISTOGRAM}(\{y_{ji}^{1:k}\}, \{w_{ji}^{1:k}\}, \text{bins} = \mathbb{C}) \quad \text{where,} \quad C \in \mathbb{C}. \tag{10}$$

The final class label estimate y_i^* for that object O_i is obtained by weighted max-pooling over all its class label estimates $\{y_{ji}^* \ \forall j = \{1:N_S\}, j \neq i\}$,

$$y_i^* = \arg\max_C \ \text{HISTOGRAM}(\{y_{ji}^*\}, \{w_{ji}\}, \text{bins} = \mathbb{C}) \quad \text{where,} \ C \in \mathbb{C} \text{ and, for } i \in \{1:N_S\}. \tag{11}$$

This process is repeated for all other objects and the best label sequence estimate for a novel scene Y^* (§IV) is obtained.

The following interpretation is the fundamental crux of our methods: A IFRC vector computed for a trajector wrt. a landmark can be portrayed as a description of the trajector from the point of view of the landmark. If we pool such incoming IFRC vectors from all objects in the scene as the landmarks, to a particular object of interest as the trajector, we can understand it as a 'description of that object by the environment'. We believe that such a portraiture of each object, by its environment is discriminative and these can be used for object classification, eliminating the need of object-centric appearance based cues.

V. EXPERIMENTS

We design our experiments involving joint object classification. Our system achieves two important aspects: (1) labelling a group of underlying structures by *considering*

them together (2) and to do so, solely using underlying structures between objects in the scene.

We use standard ball-tree implementations for KNN and RBNN and the query time is not comparable to the inference part. The time order of complexity for joint second-order treatment inference is $O(N^2K^N)$ and for approximate second-order treatment inference is $O(N^2K^2k)$ where N is number of objects in a scene, K is the number of object classes and k the number of neighbours queried. The parameters for KNN and RBNN were fixed using exhaustive grid search.

A. Dataset

We run our experiments on the *KTH-3D-Total* dataset [45] comprised of long-term office desktop observations and manually annotated objects. Desktop data can be considered as prototypical environments because they (1) are controlled, easily and repeatedly observable environments, (2) offer inter-object spatio-temporal dynamics that offer modelling predicaments, (3) exhibit a majority of real-world perception challenges: occlusions, lighting inconsistencies, photo-unfriendly objects, instance variations and intra-object dynamics (eg. Laptop open and closed).

The dataset contains 461 scenes (desktop instances) belonging to 20 desktop-users and containing 18 object classes. We used only those object classes which had 100 or more occurrences: *Monitor, Papers, Keyboard, Mouse, Mug, Lamp, Laptop, Book, Pen, PenStand, Bottle, Headphones*. To introduce controlled noise, we also consider five more categories which had more than 50 occurrences: *Folder, Mobile, Glass, Flask, Jug*; We require at least 100 occurrences to have adequate data for drawing credible conclusions. In every scene, there are 8-10 objects and there could be upto 3 instances of a particular object class.

B. Experiment Details

During experimentation, for each test run instance or *trial*, we sample from all relevant scenes in the dataset to make the test data and retain the exclusive rest of the dataset as training data in the proportion 20:80. Since our CCS is a monolithic classifier for multiclass classification, the primary challenge is perplexity caused by increased number of classes. If we consider more number of classes, we have a higher chance of confusion but at the same time, by design, we expect our system to find more evidence to ascertain each object label – making this an interesting trade off. We choose object classes in decreasing order of their frequency of occurrence; we experiment by initially considering a subset of the top 3 classes and then repeating experiments by including the next top class every time. For every such class subset, we conduct at least 10 trials to report average accuracies. When sampling test data, we prefer only those scenes containing the entire class subset to uphold *maximum perplexity*.

We compute system accuracy over one trial by averaging over class specific true positive rates, for all object classes considered in that trial. This measures how accurate our system is at a focussed, object level classification. Another measure of performance is what we call *hit@*n. It is a

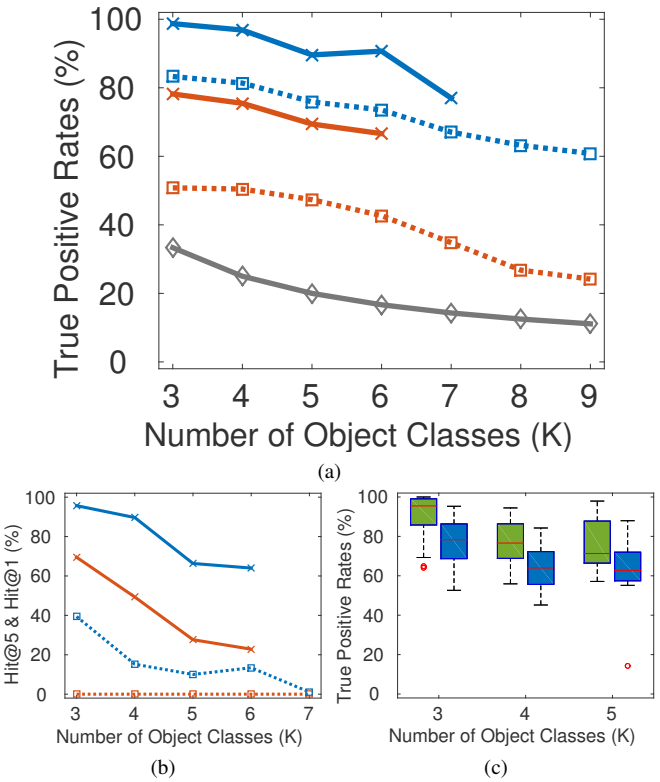


Fig. 6: Joint object classification (a) true positive rates and (b) hit@5 of different classifiers vs. number of object classes. Solid lines correspond to joint second-order and dotted lines to approximate second-order treatments respectively. Blue (–) is weighted KNN, red (–) is RBNN, grey (–) is a baseline random classifier. (c) Standard quartile statistics for accuracies of joint second-order(■) and approximate second-order(■) KNN classifiers, over 30 randomly sampled class subsets.

rougher version of the standard precision@n. A hit@n only checks if precision@n $>$ 0. Getting a high hit@n score measures if the system determined labels of all objects in a scene, although with an allowance of n guesses.

We finally evaluate the robustness of our CCS in the face of simulated realistic noise: We include object classes, other than those in the testing subset, to compose mild and high amounts of confusing clutter usually caused due to false detections. In a second test, we inject measurement noise in object pose and size and inspect the resilience of our system to such errors.

VI. RESULTS AND DISCUSSION

Clean Experiments: We tested our context comprehension system with the dataset organised according to §V-B and the accuracies are recorded in Fig.6. As the size of our class subset increases, perplexity increases, yet we observe a graceful decline in the average true positive rates. We relax the maximum perplexity constraint (§V-B) as our class subset size increases beyond 7 classes. The data is insufficient to report results for experiments containing more than 9 classes.

It is clear in Fig.6 that the more thorough joint second-order classifiers exploit their higher constrained inference to perform better than their approximate counterparts. However, the approximate second-order treatments (RBNN,KNN), averaged over number of classes (40%, 72%), already perform

comparably to the systems in [7]: (57%, 76%).

The statistics of run time per scene of the approximate treatment inference, on a standard desktop computer, over different class subset sizes and scene sizes are: mean run time is 4.3ms with a standard deviation of 2.1ms. The methods in [7] use inference schemes with order of complexity comparable to our joint treatment inference, which has a mean run time of 5×10^4ms, varying exponentially as K^N. This shows that the approximate second-order treatment gives a huge speed boost for not a drastic drop in accuracy.

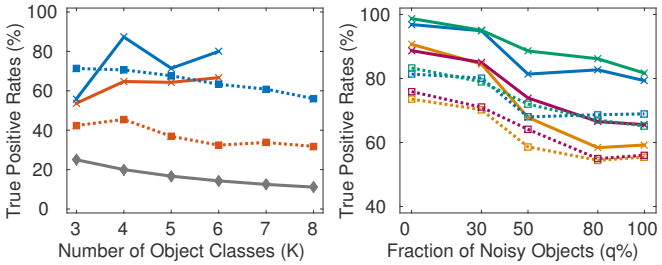
For this data, KNN classifiers significantly outperform the RBNN classifiers as shown in Fig.6a. We also expected our RBNN classifiers to provide more reliable labels because they 'listen to' more voting members, when available. However, our RBNN classifiers with fixed radii, drag in a high number of confounding data points from the neighbourhood leading to significantly diminished accuracies (Fig. 6a). Contrastingly, KNN classifiers have a constant cardinality neighbourhood (k) resulting from a non-linear truncation in the data subspace of I^7. If data is scattered such that there are crude clusters for each class, then using RBNN is advantageous; Otherwise, KNN with a tuned k, will keep out the majority of the confounding neighbours to provide better consensus.

In Fig.6b we observe the hit@5 rates for the joint second-order treatment is significantly larger than the hit@1 rates of their corresponding approximate treatments. Due to their inference process, we can only measure a hit@1 rate for the approximate second-order treatments. This means that, even though the approximate second-order treatments can provide quick reasonable true positive rates on the whole, it is better to use the joint second-order treatments when the situation demands true joint object classification.

When we test class subsets by randomly selecting classes, our systems still perform steadily and consistently as shown in Fig. 6c. We also experimented by excluding scenes of those desktop-users, selected in the test set, from the training set. This ensured that our system generalises learnt spatial relation structures across scene types and not do a template matching in Ξ. The accuracies of such experiments matched those values in Fig.6 within $\pm 2\%$.

Noise Experiments: Our CCS performs as shown in Fig.7a when we introduce detection noise by creating an exclusive *ghost class* out of 3 component classes exclusive to the class subset in question. Increasing the amount and spread of ghost class data in the data space with more component classes makes it harder for multiclass classification. Our CCS refrains from confusion and has a high true positive rate for all members in the class subset and also the ghost class. The joint second-order treatments perform as well as the best approximate second-order treatments (KNN) and have a slight increasing trend in Fig.7a because larger class subsets provide more evidence to constrain inference toward more accurate results. The drop in accuracy at the start is because |class subset|:|ghost class| is 3:3 causing a lower signal-to-noise-ratio in comparison to the 4:3, 5:3,... cases. We increased ghost class size to 5 and the accuracies of the classifiers do not vary significantly.

The results in Fig.7b shows the decline of accuracies



(a) (b)

Fig. 7: Robustness to noise: (a) *Detection Noise*: 'ghost' class is made using 3 classes exclusive of the testing classes subset. Plot color codes are consistent with Fig. 6. (b) *Measurement Noise*: Solid and dotted lines correspond to KNN joint second-order and KNN approximate second-order classifiers. These classifiers are tested by considering 3■, 4■, 5■, 6■ object classes during testing (Note the Y-axis limits).

of joint second-order and approximate second-order KNN classifiers with increase in measurement noise and class subset size. The measurement noise is injected for a fraction q of the objects in every scene, randomly altering: object size by upto $\pm 25\%$, position by upto $\pm 25\%$ of object's dimensions projected along the axes and orientation by upto $\pm 45^\circ$ in yaw angle. Consider, the curve corresponding to class subset of size 5; The joint second-order treatment KNN continues to give an accuracy of 67%. This experiment is our most important, showing that our CCS is significantly robust to everyday, detection-measurement problems.

Even though our CCS is robust, less complex, faster and provides opportunity for online knowledge growth, there are still some challenges. Non-parametric methods are sensitive to r value (RBNN) and k value (KNN) tuning, and the metric used [46]. We can carefully calibrate these parameters by grid searching for optimum performance once the data domain is fixed. The data dependence of the lazy learners makes it susceptible to outliers, confusing data scatter and data bias. We can circumvent these problems by correct weighting and using reinforcement learning schemes in an online setting.

Deep learning approaches for scene understanding–as in [47], [48], [49]–require numerous, class specific, ground truth examples (≈3000 samples per class) and intensive training over couple of hours [50]. Our CCS learns and generalises almost instantaneously, with smaller training data (≈30 samples per class). The training of our CCS is inexpensive, conducive for easy online learning and HRI friendly.

VII. CONCLUSIONS AND FUTURE WORK

In this work, we have developed a context comprehension system to aid robotic vision for scene understanding through joint object classification. We introduce Intrinsic Frame of Reference Calculi – a HRI conducive, spatial relation feature set for structure learning which is highlighted by its robustness to corrupt object detections. Our context comprehension system operates on IFRC with a more accurate joint second-order and faster approximate second-order treatments. The context intelligence systems employ non-parametric methods which provide for on-the-fly learned knowledge modification and faster operation. By experimenting with our systems

on desktop data, we were able to show that we captured sufficient information from only structures in scenes to conduct joint object classification not only to aid but sometimes perform even as good as state-of- the-art appearance based perception systems! When compared to prevalent systems [7], The joint second-order treatment provides significantly better accuracies at similar complexities; The approximate second-order treatment is as accurate and more robust with a huge improvement in speed of reasoning.

In the immediate future, we will modify the joint second-order inference system to work in real time and integrate with an appearance based perception system on our robotic platform. We want to improve knowledge and inference by implementing reinforcement learning to rank learnt exemplars based on contribution history to classification accuracy. Subsequently, we wish to extend this system to act as a generative system to help in object search and anomaly detections.

ACKNOWLEDGMENT

We thank Asst.Prof. Carl Henrik Ek for crucial discussions; Alejandro Marzinotto, Johan Ekekrantz, Prof. Thippur Sreenivas for support discussions. This research has been funded by the EU Project STRANDS (ICT-2011-600623).

REFERENCES

[1] T. Fulhammer *et al.*, "Autonomous learning of object models on a mobile robot," *Robotics and Automation Letters*, pp. 26–33, 2017.

[2] Z. Teng and J. Xiao, "Surface-based detection and 6-dof pose estimation of 3-d objects in cluttered scenes," *Transactions on Robotics*, pp. 1347–1361, 2016.

[3] S. Gupta *et al.*, "Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation," *International Journal of Computer Vision*, pp. 133–149, 2015.

[4] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European Conference on Computer Vision*, 2014, pp. 345–360.

[5] N. Silberman *et al.*, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*, 2012, pp. 746–760.

[6] J. Shotton *et al.*, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *European conference on computer vision*, 2006, pp. 1–15.

[7] A. Thippur *et al.*, "A comparison of qualitative and metric spatial relation models for scene understanding." in *AAAI*, 2015, pp. 1632–1640.

[8] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: A survey," *Foundations and Trends in Computer Graphics and Vision*, pp. 177–280, 2008.

[9] A. Thippur, C. H. Ek, and H. Kjellstrm, "Inferring hand pose: A comparative study of visual shape features," in *Automatic Face and Gesture Recognition*, 2013, pp. 1–8.

[10] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends in Cognitive Sciences*, pp. 520 – 527, 2007.

[11] G. Heitz and D. Koller, *Learning Spatial Context: Using Stuff to Find Things*, 2008, pp. 30–43.

[12] M. Bar, "Visual objects in context," *Nature Reviews Neuroscience*, pp. 617–629, 2004.

[13] M. Marszalek *et al.*, "Actions in context," in *Computer Vision and Pattern Recognition*, 2009, pp. 2929–2936.

[14] K. Dubba *et al.*, "Interleaved inductive-abductive reasoning for learning complex event models," in *Conference on Inductive Logic Programming*, 2011, pp. 113–129.

[15] H. A.Pieropan, C.H.Ek, "Recognizing object affordances in terms of spatio-temporal object-object relationships," in *International Conference on Humanoid Robots*, 2014.

[16] A. Torralba, "Contextual priming for object detection," *International Journal of Computer Vision*, pp. 169–191, 2003.

[17] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Computer Vision and Pattern Recognition*, 2009, pp. 413–420.

[18] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in Neural Information Processing Systems*, 2011, pp. 109–117.

[19] B. Fulkerson *et al.*, "Class segmentation and object localization with superpixel neighborhoods," in *International Conference on Computer Vision*, 2009, pp. 670–677.

[20] L. Ladicky *et al.*, *Graph Cut Based Inference with Co-occurrence Statistics*, 2010, pp. 239–253.

[21] C. Galleguillos *et al.*, "Object categorization using co-occurrence, location and appearance," in *Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[22] J. Chen *et al.*, "A survey of qualitative spatial representations," *The Knowledge Engineering Review*, pp. 106–136, 2015.

[23] Y. Gatsoulis *et al.*, "Qsrlib: a software library for online acquisition of qualitative spatial relations from video," in *Workshop on Qualitative Reasoning at IJCAI*, 2016.

[24] T. o. Krajnik, "Spectral analysis for long-term robotic mapping," in *IEEE International Conference on Robotics and Automation*, 2014, pp. 3706–3711.

[25] P. Duckworth *et al.*, "Unsupervised learning of qualitative motion behaviours by a mobile robot," in *International Conference on Autonomous Agents*, 2016, pp. 1043–1051.

[26] B. Rosman and S. Ramamoorthy, "Learning spatial relationships between objects," *International Journal of Robotics Research*, pp. 1328–1342, 2011.

[27] H. M. Dee *et al.*, *Scene Modelling and Classification Using Learned Spatial Relations*, 2009, pp. 295–311.

[28] J. M. Keller and X. Wang, "Learning spatial relationships in computer vision," in *International Fuzzy Systems*, 1996, pp. 118–124 vol.1.

[29] J. Freeman, "The modelling of spatial relations," *Computer Graphics and Image Processing*, pp. 156 – 171, 1975.

[30] A. Gupta *et al.*, *Blocks World Revisited: Image Understanding Using Qualitative Geometry and Mechanics*, 2010, pp. 482–496.

[31] W. Choi *et al.*, "Understanding indoor scenes using 3d geometric phrases," in *Computer Vision and Pattern Recognition*, 2013, pp. 33–40.

[32] D. Lin, S. Fidler, and R. Urtasun, "Holistic scene understanding for 3d object detection with rgbd cameras," in *International Conference on Computer Vision*, 2013, pp. 1417–1424.

[33] M. Tenorth and M. Beetz, "Knowrob: A knowledge processing infrastructure for cognition-enabled robots," *Journal of Robotics Research*, pp. 566–590, 2013.

[34] M. Skubic *et al.*, "Spatial language for human-robot dialogs," *IEEE Transactions on Systems, Man, and Cybernetics*, pp. 154–167, 2004.

[35] M. Aly, "Survey on multiclass classification methods," *Neural Networks*, pp. 1–9, 2005.

[36] A. Elgammal *et al.*, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proceedings of the IEEE*, pp. 1151–1163, 2002.

[37] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, pp. 207–244, 2009.

[38] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern recognition*, pp. 2038–2048, 2007.

[39] T.-H. Chiang *et al.*, "A ranking-based knn approach for multi-label classification." *ACML*, pp. 81–96, 2012.

[40] L. Kunze *et al.*, "Combining top-down spatial reasoning and bottom-up object class recognition for scene understanding," in *Intelligent Robots and Systems*, 2014, pp. 2910–2915.

[41] J. Young *et al.*, "Towards Lifelong Object Learning by Integrating Situated Robot Perception and Semantic Web Mining," in *European Conference on Artificial Intelligence*, 2016.

[42] J. H. Bappy *et al.*, *Online Adaptation for Joint Scene and Object Classification*, 2016, pp. 227–243.

[43] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern Recognition*, pp. 2038 – 2048, 2007.

[44] F. H. Previc, "The neuropsychology of 3-d space." *Psychological bulletin*, p. 123, 1998.

[45] A. Thippur *et al.*, "Kth-3d-total: A 3d dataset for discovering spatial structures for long-term autonomous learning," in *International Conference on Control Automation Robotics Vision*, 2014, pp. 1528–1535.

[46] M. R. Abbasifard *et al.*, "A survey on nearest neighbor search methods," *International Journal of Computer Applications*, pp. 39–52, 2014.

[47] R. Socher *et al.*, "Convolutional-recursive deep learning for 3d object classification." in *NIPS*, 2012.

[48] T. Malisiewicz and A. Efros, "Beyond categories: The visual memex model for reasoning about object relationships," in *Advances in Neural Information Processing Systems 22*, 2009, pp. 1222–1230.
[49] J. Sun and D. W. Jacobs, "Seeing what is not there: Learning context to determine where objects are missing," *arXiv preprint*, 2017.
[50] H. Azizpour *et al.*, "From generic to specific deep representations for visual recognition," *CoRR*, 2014.