Depth-informed Qualitative Spatial Representations for Object Affordance Prediction

Alexia Toumpa[1] and **Anthony G. Cohn**[2, 3, 4, 1]

Abstract. Acquiring knowledge about object interactions and affordances can facilitate scene understanding and human-robot collaboration tasks. As humans tend to use objects in many different ways depending on the scene and the objects' availability, learning object affordances in everyday-life scenarios is a challenging task particularly in the presence of an open set of interactions and objects.

We address the problem of affordance prediction for class-agnostic objects with an open set of interactions; we achieve this by learning similarities between object interactions in an unsupervised way and thus inducing clusters of object affordances. A novel depth-informed qualitative spatial representation is proposed for the construction of the *Activity Graphs* (AGs) which abstract from the continuous representation of spatio-temporal interactions in RGB-D videos. These AGs are clustered to obtain groups of similar object affordances. Our experiments in a real-world scenario do not impose any object or scene constraints and demonstrate that our method handles object occlusions and learns object affordance clusters with a high V-measure.

1 INTRODUCTION

In the literature, the meaning of the term *affordance* of an object differs depending on the context. In robotic applications, *e.g.* robot manipulation tasks, the definition of *affordance* is bound to the part of a tool which can be afforded in a specific way, *e.g.* the handle of a hammer has the affordance of 'hold' whereas the head has the affordance of 'hit'. In contrast, in human-object interaction recognition tasks, *affordance* is defined as the way an object can be utilized by the human in a scene, *e.g.* if a human uses a cup for containing something then the cup will have the affordance of 'contain'. Moreover, any object may have more than one affordance as it depends on the purpose it is being used for, *e.g.* a pizza box can have the affordance of 'contain' when it is being utilized as a container of a pizza or 'support' when is plays the role of a tray, and such multi-labelled affording objects can be recognized by considering their interactions with other objects.

In a human-robot collaboration scenario, acquiring knowledge of the affordances of the objects in a scene is crucial for aiding the human, *e.g.* assisting the human when performing a physically hard task. This becomes challenging when the scene comprises an open-set of arbitrary objects and the affordance space enlarges. Moreover, in human action prediction tasks, the affordances of the objects carry useful information for the prediction of the future action, and are highly correlated to the rest of the objects the human interacts with. Nevertheless, such knowledge is not easy to obtain as humans tend to use the same object in different ways depending on the performed task, thus changing its primary affordance.

Object *affordances* were first formally defined by J. Gibson [10]; however the concept of object affordances as it is understood in Computer Science has not been explicitly defined in the literature causing some confusion. For this purpose we propose the definition:

Definition 1 *An affordance is a property of an object arising from its interaction with another entity, i.e. agent, object. It is correlated to the occurring interaction as every interaction exploits at least one object affordance. An object may have multiple affordances based on the various interactions it may have with other entities.*

Based on this definition, and building on our previous work [30], we address the problem of affordance prediction by exploiting pair-wise object interactions through RGB-D video data. Hence, we rely on the way objects are being utilized by the human agents in a scene, allowing objects to support different kind of affordances at the same time. We focus on learning groups of high-level object interactions which take into account their spatio-temporal relations from extracted visual appearances. Graphs are able to capture high level information of relationships or even dynamic relational changes. Thus, we represent these pair-wise object interactions through a high-level graphical structure, the *Activity Graph* (AG), while abstracting from the continuous spatio-temporal representation and acquiring depth-informed qualitative spatial relations between object pairs.

Definition 1 states that *"every interaction exploits at least one object affordance"*; affordances of objects are inferred from these high-level graphs representing pair-wise object interactions. Affordance clusters are formed in an unsupervised way by exploiting intra-class graph similarity using a *set Edit Distance* (sED) measure. By clustering graph structures, a hierarchical tree representation is produced demonstrating their similarity. Since our approach is based on learning a high-level representation of interactions, it is not limited to any number or kind of affordances, scenes, and objects.

To obtain a richer set of spatial relationships than those possible from a sequence of purely 2D frames, we exploit the depth information assuming the presence of RGB-D video data. The depth cues allow some inference about the morphology of the objects in the scene and thus the way they can interact with other objects, *e.g.* 'concave' objects can act as *containers*.

Hence, the objectives of this work are:

- to propose a depth-informed set of qualitative relations which can describe a wide variety of object configurations regarding a real-world scenario as well as handling object occlusions, for detecting effectively spatial relationships of object interactions

[1] University of Leeds, UK, email: {scat,a.g.cohn}@leeds.ac.uk
[2] Luzhong Institution of Safety, Environmental Protection Engineering and Materials, and School of Mechanical and Electrical Engineering, Qingdao University of Science and Technology, China
[3] Turing Fellow at the Alan Turing Institute, UK
[4] Department of Computer Science and Technology, Tongji University, China

- to capture representations of object interactions from multiple object affordances
- to evaluate our method on a challenging dataset which comprises real-world scenario scenes with object occlusions.

2 RELATED WORK

Several methods have been proposed for detecting functional object parts and their corresponding affordance labels. These works involve the detection of object affordance parts by considering their visual characteristics and their geometric features. One of the early works in this direction focused in the detection of graspable object areas by creating local visual descriptors of grasping points and estimating the probability of the presence of a graspable object based on the Bernoulli trial [19]. New approaches employ Convolutional Neural Network (CNN) models to produce classes of functional object parts from RGB data [22, 7, 27]. However, depth cues along with the RGB information have demonstrated a greater detection accuracy in this task [21, 20]. Additionally, incorporating knowledge about the scene and context in which an object is being used boosts even more the prediction accuracy of such system [33].

However, processing static visual information restricts the number of affordances assigned to an object to be the ones correlated only with its visual features. For this purpose, many works have considered exploiting the correlation of human actions and the detected objects in a scene [11, 32, 8]. Depending on the human-object interaction being held, a different affordance is detected. These works demonstrate that by fusing knowledge about the scenario in which an interaction takes place enforces the prediction of affordances, however limits the generalizability across different domains.

To accommodate domain independence, high-level graph representations of interactions are being employed. Recent works introduce such graphical structures in synthetic indoor environments focusing on the prediction of furniture areas the human is most likely to interact with [25], whereas outdoors scenes are examined by considering the behavior of moving objects around them [31]. Nevertheless, these approaches only consider a one-to-one mapping of affordances and objects, hence the objects are bound to a single kind of interaction, not permitting multi-labelled object affordances.

Though graphical structures are able to retain high-level information about the interactions in the scene, their structure might be the cause of domain restriction [2, 1, 24, 23]. For this purpose, qualitative spatio-temporal relations are exploited for their construction [28].

One of the fundamental obstacles in these works is object occlusion. To mitigate this problem, the tracks and visual appearances and disappearances of non-deformable objects are considered [16, 15, 18]. However, these approaches are restricted to the detection of a single object affordance, *i.e.* containment, and no consideration of non-containment relation due to occlusion is handled.

In this work, we propose a novel method for predicting object affordances in real-world scenarios in the presence of object occlusions. Different from any other published work, our approach is not restrained to a predefined set of objects, interactions, or scenes. High-level graphs assist in considering an open-set of interactions and no object labels restrict the generalizability of our method. Our experiments demonstrate how, without any supervision, we acquire homogeneous and complete clusters of object affordances by exploiting qualitative information about their interactions and their shapes.

3 BACKGROUND

Relational graph structures represent high-level information by abstracting from the continuous space of the exploited relations of the graph entities. From definition 1: *"It (an affordance) is correlated to the occurring interaction as every interaction exploits at least one object affordance."*, hence an object-object interaction reveals each object's affordances. Relational graph structures of pair-wise object interactions aid in representing an open set of interactions. For this purpose, in this work we exploit the representation of *Activity Graphs* (AGs) [29] which are relational graphs that describe the interaction between entities, *i.e.* objects, of a scene by considering their qualitative spatio-temporal relationships.

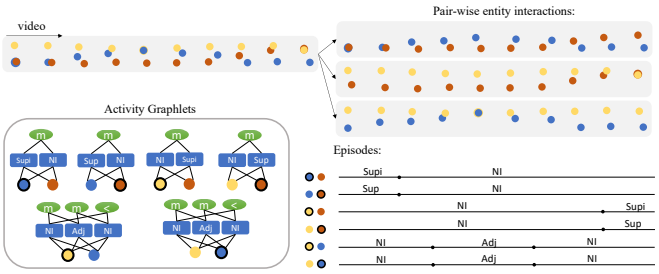

Figure 1. Qualitative spatial relations are extracted from detected episodes in the temporal domain of a video. *Activity Graphlets* are constructed using these qualitative spatio-temporal relations for individual detected objects (encircled) in the scene describing their interaction with another object.

An AG consists of three layers of vertices, where each layer comprises a single type of nodes and only nodes in adjacent layers can be connected with each other. The three layers of an AG are: the *object*, the *spatial*, and the *temporal* layers. The object layer contains the set of vertices of the objects or entities which interact (V_{obj}), the spatial layer consist of vertices with the spatial relations (V_{spat}) which describe the spatial interactions of the entities in V_{obj}, and the temporal information of the occurrences between the spatial relations exists in the vertices of the temporal layer (V_{temp}).

Qualitative *DiSR* spatial relations, introduced in Section 5, are employed to describe the spatial relationships of objects (V_{spat}), whilst *Allen's temporal algebra* [3] is exploited to express the temporal relationships between the spatial relations (V_{temp}). This qualitative temporal set consist of the relations: 'before' ($<$, $>$), 'meets' (m, mi), 'overlaps', (o, oi), 'starts' (s, si), 'during' (d, di), 'finishes' (f, fi), and 'equals' ($=$). The set of qualitative spatial relations for every interaction is obtained from the presence of *episodes*. An *episode* is defined as the maximum period of time where a single spatial relation between two entities holds, while a different spatial relation occurs before and after the defined time period. By considering *episodes* instead of individual frames for detecting the spatial relationships of the objects, our method can effectively generalize across different video fps and is not video length dependent.

As many affordances are revealed by the pair-wise interactions of the objects, we define an *Activity Graphlet* ($AGraphlet$) as a subgraph of an AG which carries the spatial and temporal information ($V^{'}_{spat}$, $V^{'}_{temp}$) of every pair of objects ($V^{'}_{obj}$) in a video scene (Fig. 1). Every graph describes an object's interaction, hence two interacting objects have inverse graphs assigned to them.

4 OVERVIEW

The proposed approach exploits depth-informed qualitative information to solve the problem of object affordance prediction. Our method is based on unsupervised learning of detected object interactions. A hierarchical clustering is employed for this purpose, described in Section 6, exploiting high-level interaction representations and producing a dendrogram of object affordances. $AGraphlet$s are used

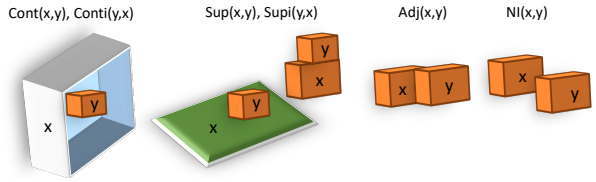

Figure 2. *DiSR* representations in 3D space.

as a high-level representation of object interactions, thus achieving a generalization across various scenarios. To effectively represent qualitative spatial relationships between objects we describe in Section 5 a novel set of depth-informed spatial relationships which infers more accurate object relative relations. Such qualitative spatial information constitutes the spatial layer of an $AGraphlet$.

Objects of interest are localized in the input frames of a video by utilizing any visual-based object detector algorithm which provides class-agnostic object bounding boxes.

5 DEPTH-INFORMED RELATIONS

Though simple and discrete spatial interactions, *e.g.* 'touching' and 'not touching' in 2D space, are captured effectively in the 2D image plane, the determination of more complex spatial relationships, *e.g.* 'supporting', 'containing', is challenging, especially when considering a cluttered scene. To address this limitation we propose a depth-informed set of qualitative spatial relationships which take into account the object's convexity-type. The object's convexity-type carries information about the object's affordance, hence the kind of interaction which can hold between that object and another. For example, a concave object, due to its concavity, can afford the relation of *contain* by playing the role of the *container*, whilst an object interacting in a specific way with an object which has the affordance *container*, can be regarded as a *containee*.

The spatial information employed to describe the object interactions in the spatial domain, consists of a discrete set of spatial relations that can describe effectively any spatial configuration of the objects in a scene. These depth-informed interactions are exploited for the construction of more accurate qualitative graphical structures, achieving more homogeneous and complete affordance clusters.

5.1 Formulation of *DiSR*

We propose the set of *Depth-informed Spatial Relations* (*DiSR*) (Fig. 2): 'supports' (Sup, Supi), 'contains' (Cont, Conti), 'adjacent' (Adj), 'not interacting'(NI). The proposed set takes into account the 2.5D information of the detected objects rather than only the 2D projections in the camera plane as in the primary definition of the RCC [26, 6]. The depth information is considered for reasoning about the spatial relative positions of objects in 3D space, hence enabling the distinction between occlusions and interactions. Furthermore, from the depth cues we can effectively infer the convexity-type of the detected objects, and along with the corresponding depth differences, *DiSR* relations are extracted. For a spatial interaction to hold, a depth distribution overlap must be evident, and depending on the convexity-type of the interacting objects and in which part of the distribution the overlap appears, a different qualitative spatial relation from the set of *DiSR* holds.

DiSR spatial relations take into account the depth information of the detected object masks in the scene as well as their 2D location considering the bounding box enclosing the detected mask of every object. We present in Table 1 the definition of every *DiSR* relation inspired by the RCC set, whilst exploiting the RCC relations: 'overlap' (O) and 'part' (P, Pi) computed in the 2D image plane.

Table 1. DiSR

DiSR	Definition	Description
Sup(x,y)	$(\mathsf{DPO}(x,y) \wedge \text{surface}(x)) \vee \mathsf{On}(y,x)$	x supports y
Cont(x,y)	$\mathsf{P}(y,x) \wedge \text{concave}(x) \wedge \mathsf{DPP}(y,x)$	x contains y
Adj(x,y)	$\mathsf{O}(x,y) \wedge \mathsf{DPO}(x,y) \wedge \neg\mathsf{Cont}(x,y) \wedge \neg\mathsf{Cont}(y,x) \wedge \neg\mathsf{Sup}(x,y) \wedge \neg\mathsf{Sup}(y,x)$	x is adjacent to y
NI(x,y)	$\neg\mathsf{Sup}(x,y) \wedge \neg\mathsf{Sup}(y,x) \wedge \neg\mathsf{Cont}(x,y) \wedge \neg\mathsf{Cont}(y,x) \wedge \neg\mathsf{Adj}(x,y) \wedge \neg\mathsf{Adj}(y,x)$	x does not interact with y

The convexity type of an object, being either *concave*, *surface* or *convex*, is described further in Section 5.2. 'Depth Overlap' (DPO) and 'Depth Proper Part' (DPP, DPPi) are primitive relations which hold between objects' depth distributions, defined as:

$$\mathsf{DPO}(x,y) \equiv ((dmax_x \geq dmin_y) \wedge (dmax_x < dmax_y) \wedge (dmin_x < dmin_y)) \vee ((dmax_y \geq dmin_x) \wedge (dmax_y < dmax_x) \wedge (dmin_y < dmin_x))$$

$$\mathsf{DPP}(x,y) \equiv (dmax_x > d_cmin_y) \wedge (dmax_x \leq d_cmax_y) \wedge (dmin_x \geq d_cmin_y) \wedge (dmin_x < d_cmax_y)$$

where *dmax* and *dmin* are the maximum and minimum depth values, respectively, by considering the depth cues of the detected object's mask, and d_cmax and d_cmin are derived from Alg. 2.

Moreover, the spatial relation On is defined as,

$$\mathsf{On}(x,y) \equiv \mathsf{O}(x,y) \wedge ((ymax_x \geq ymax_y) \wedge (ymin_x \geq ymin_y) \wedge (xmax_x \leq xmax_y) \wedge (xmin_x \geq xmin_y))$$

where $(xmin, ymax)$ and $(xmax, ymin)$ are the top-left and bottom-right corners of the detected object's bounding box, respectively.

The definitions proposed are an approximation of the English meaning they are referring to; however it is possible to satisfy the definitions by configurations which do accord with intuitive meaning of the English word. *E.g.* consider the case where three objects are stacked the one on top of the other, then in the 2D image plane On(*top object*, *bottom object*) will be True even though there is a *middle object* in between. Nevertheless, these definitions are easy to compute and work well in the everyday scenes we have considered so far. In future work we may refine these relationships to more closely correspond to the semantics of the English words.

5.2 Object's Convexity-type

We employ Alg. 1 to determine every object's convexity type, by considering its depth distribution, extracted from the depth cues. Detected objects are grouped into three convexity-type categories: *convex*, *concave* or *surface*, in reference to their depth distribution. The proposed algorithm is based on a convexity depth threshold ($thresh_{convex}$) which defines the upper boundary of depth range information of a 'convex' type object, the selection of which was determined from an empirical study. Objects with depth range greater than $thresh_{convex}$ are subject to be grouped under 'concave' or 'surface' depending on their depth contour hierarchies (*ContourHierarchy*). A depth contour is the contour created by the depth information exceeding $thresh_{convex}$. Contour hierarchies establish a tree structure of contour inclusion, where every node of the tree stands for a contour and every parent includes its children. Thus, the detection of a child con-

Algorithm 1 Define the convexity type of an object.

Given: $thresh_{convex}$

procedure OBJECTCONVEXITY($\text{dist}_{\text{depth}}$)
  $dmax \leftarrow max(dist_{depth})$; $dmin \leftarrow min(dist_{depth})$
  if ($dmax - dmin$) $< thresh_{convex}$ **then**
    $object_{type} \leftarrow$ convex
  else
    $C \leftarrow$ ContourHierarchy($dist_{depth}$)
    if $C.child()$ exists **then**
      $object_{type} \leftarrow$ concave
    else $object_{type} \leftarrow$ surface
  return $object_{type}$
  end

Algorithm 2 Define min and max depth of an object's concavity.

Given: h, n

procedure CONVEXITYDEPTH($\text{dist}_{\text{depth}}$)
  $dmax \leftarrow max(dist_{depth})$; $dmin \leftarrow min(dist_{depth})$
  $object_{type} \leftarrow$ OBJECTCONVEXITY($dist_{depth}$)
  if $object_{type} =$ concave **then**
    $sections \leftarrow$ ($dmax - dmin$)$/h$
    $d_cmax \leftarrow dmax$
    $d_cmin \leftarrow dmax - (n * sections)$
  else $d_cmax \leftarrow dmax$; $d_cmin \leftarrow dmin$
  return d_cmax , d_cmin
  end

tour in the depth domain, deduces the presence of a concave curve, therefore a 'concave' type object, and 'surface' type otherwise.

5.3 Depth of Convex and Concave Objects

By estimating the distribution of the depth information we obtain knowledge about the *indentation area* (m-) and *protrusion area* (M+) of an object, as defined in the *Process-Grammar* [14]. More specifically, visually 'convex' type objects do not appear to have any *indentation areas* whereas 'concave' type objects are characterized by their concavity curve which morphologically appears as a bay-formation described as M+m-M+. Such information is critical to ascertain a relation when two objects interact in the 2D image plane, *e.g.* a Cont *DiSR* relation occurs between one or more 'concave' type objects when the depth information of the *containee* confirms that is between m- and M+ areas of the *container*. We propose Alg. 2 to infer the boundaries of the m- and M+ areas, as a direct generalization of the *Process-Grammar* to 2.5D, with respect to the object's depth information and convexity-type, indicating a concave curve in 3D space. For a 'concave' type object we partition the depth information into h sections for distinguishing the *indentation* from the *protrusion area*. The n sections with the highest depth values are estimated to capture its concave curve. The parametrization of h and n was conducted in an empirical study. We set the depth boundaries of such objects to enclose the concave curve's depth information for detecting the relation Cont. Depth boundaries of 'convex' and 'surface' type objects are not being processed due to concave curve absence.

6 LEARNING OBJECT AFFORDANCES

We learn object affordances by clustering $AGraphlet$s , whilst employing a hierarchical approach. Every $AGraphlet$ represents an object's interaction with another object in the scene. We consider an interaction between an object pair as the spatio-temporal sequence of relations holding during an activity. By clustering such graph structures we produce a hierarchy of similar affordances, which are closely related with the way every object is being used in an activity. Hence, our method does not pose any constraints in the number of affordance clusters an object can be assigned to, whilst every object has as many $AGraphlet$s as detected interactions. *E.g.* consider the scenario where an agent picks a bowl from a table and places it in the microwave. The clustering mechanism examines the interaction of the bowl with the table and the bowl with the microwave as two different interactions; two $AGraphlet$s will be exploited for the bowl implying that such an object is a *supportee* as well as a *containee*.

To cluster $AGraphlet$s , we measure the difference of every graph structure with every other by looking at their spatio-temporal differences. As every graph represents a pair-wise interaction, we therefore measure the difference between pairs of interactions. Let $G^{\alpha}$and G^{β}be two $AGraphlet$s , each representing an interaction of the objects α and β respectively, with some other object. The spatio-temporal difference of these graphs is measured as defined in Eq. 1.

$$V_R^{'\alpha,\beta} = \{v : v \in \{V_R^{'\alpha} \setminus V_R^{'\beta}\} \cup \{V_R^{'\beta} \setminus V_R^{'\alpha}\}\} \quad \text{where } R \in \{spat, temp\} \tag{1}$$

For clustering graph structures, we propose the *set Edit Distance* (sED) measure which captures the similarity between graph structures according to their vertex differences, shown in Eq. 2,

$$sED = c_{spat} \sum\nolimits_{v \in V_{spat}^{'\alpha,\beta}} v + c_{temp} \sum\nolimits_{v \in V_{temp}^{'\alpha,\beta}} v \tag{2}$$

where c_{spat} and c_{temp} are the two normalized weights for the spatial and temporal vertex differences, respectively, and $c_{temp} = 1 - c_{spat}$.

We exploit the sED measure to estimate the difference of every detected object interaction, *i.e.* $AGraphlet$, with every other, as well as combine all the sED values in a *distance matrix*; the rows and the columns include the set of detected object-wise interactions and every value of the *distance matrix* stands for the measured sED difference between the respective row and column interactions. A hierarchical clustering is performed on this *distance matrix*, exploiting its sED values as the difference between the clustering datapoints, and producing a hierarchy of similarities of the interactions comprising its rows and columns. Fig. 3 illustrates a subset of the complete hierarchical clustering output in the form of a dendrogram. The full dendrogram comprises 21 clusters, which are formed by grouping the leaves of the dendrogram with respect to their hierarchy. Every cluster consists of one or more $AGraphlet$s that represent the interactions of the datapoints of that cluster. Such a hierarchy reveals the similarities of the different interactions occurring in the data, which declares a set of affordance clusters.

7 EVALUATION

7.1 Experimental Setup

For the evaluation of the proposed approach we compared the performance of the clustering mechanism of affordances. Moreover, we conducted experiments on the approach of Sridhar et al. [28] which, to the best of our knowledge, is the most recent work that exploits qualitative relational graphs to capture functional object clusters.

From the video data we extract statio-temporal pair-wise object interactions from which the object affordances are revealed. We use the CAD-120 dataset [13] which comprises RGB-D video data of 10 human activities performed 3 times by 4 different actors. These activities are everyday-life scenarios, with various configurations of the objects in the scene as well as different camera orientations. The activities are: 'arranging objects', 'taking food', 'making cereal', 'stacking', 'unstacking', 'microwaving food', 'taking

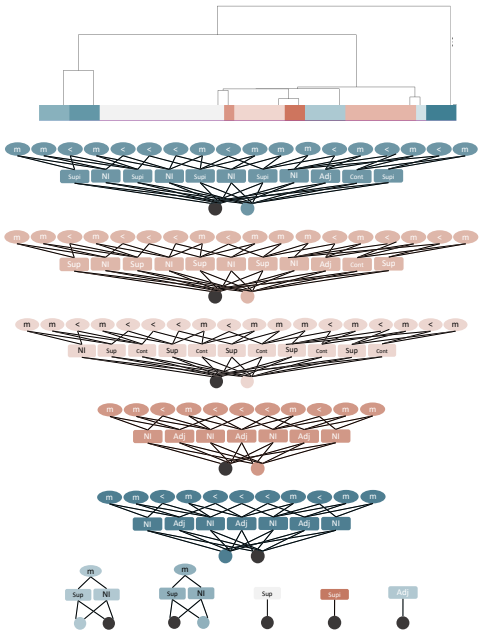

Figure 3. (best viewed in color) Dendrogram of similar object interactions from the hierarchical clustering. Color-coded graphs for each cluster represent the kind of interaction each cluster captures. Due to high graph complexity of the output, this is a sample of the full dendrogram.

medicine', 'having meal', 'cleaning objects', and 'picking objects'. Given the current representation language described in Sec. 5 of this paper, the affordances detectable in these activities are: 'container', 'containee', 'supporter', 'supportee', and 'interactive'. Our method is able to create more fine-grained groups of affordances from the groundtruth set by considering the temporal information and it is not restricted to the spatial relations occurring.

Furthermore, we employ the QSRlib library [9] for the construction of AGs. For the h and n values which are used in defining the *indentation area* of a concave object, we selected the values of 5 and 3 respectively, after conducting an empirical study for $h \in \{2,3,4,5,6,7\}$ and $n \in \{1,\ldots,h-1\}$. We also set $thresh_{convex}$ value to be 4 after evaluating on the values $1,2,3,4,5,6,7,8,9,10$,

7.1.1 *Object Detection & Tracking*

Object locations and depth information are provided from the predicted objects' masks by employing the Mask R-CNN framework [12] trained on the COCO dataset [17]. The box enclosing the object's mask corresponds to the object's bounding box. However, since objects' bounding boxes predictions are sparse, we achieve an enhancement in the object tracking system, by exploiting the CSRT tracker [5] on the Mask R-CNN predicted bounding boxes. We enrich the objects tracks by considering the CSRT predictions from the latest object bounding box occurrence, for frames where Mask R-CNN failed to detect the object. No object labels are exploited for this purpose. A threshold of minimum 0.5 IoU overlap, determined from an empirical study, is required to assign a bounding box prediction as the predicted location of a detected object.

7.1.2 *Semantic Depth Map*

The predicted object mask's depth information is employed to infer the convexity type of every object. We exclude any pixel which is part of a human detection by retrieving a human mask from the DensePose framework [4]. Whilst Mask R-CNN produces object masks for every object separately, overlapping objects have overlapping masks, thus the intersected mask area may cause problems in determining the object's convexity type. A *semantic depth map* is constructed for every frame of the video data, consisting of all the predicted object masks while eliminating any detected intersection mask area. This is achieved by assigning every pixel of such area to the object with the highest mask detection score.

7.2 Results

We used 80% of the CAD-120 dataset to determine the parameters as described above and then evaluate the proposed approach on the remaining 20% unseen videos of the dataset. We inspect the clusters of affordances formed and we evaluate their homogeneity and completeness. The datapoints being clustered consist of detected objects with their interactions with any other object, allowing multiple datapoints to point to the different interactions a single object might hold. We evaluate the clusters by reporting the normalized *v-measure*, *homogeneity*, *completeness*, and *normalized mutual information* (*NMI*) scores, with higher values implying a better clustering.

To evaluate our clustering mechanism we perform a comparison between our approach and several baselines:

- *A1C*: for evaluating the *homogeneity* score, we create a single cluster containing all the datapoints, thus achieving a *completeness* score of 1.0.
- *MC*: for evaluating the *completeness* score, we cluster every datapoint in a different cluster, obtaining a *homogeneity* score of 1.0.
- *Oracle*: we examine the best *v-measure* performance by inspecting the clustered data and the groundtruth labels and creating the optimal set of clusters. Such baseline denotes the upper performance limit of every experimental set.

We perform two kinds of experiments to investigate how well the clustering mechanism can generalize across different data samples. Both experiments comprise the test set and have the same proportion of groundtruth affordance labels, which matches the one of the whole dataset. The first experiment consists of video data with activities presenting the same spatial interactions however in inverse detection order, *e.g.* putting an object in a container and retrieving an object from a container are inverse activities. The second includes very diverse activities in the temporal as well as in the spatial relational domain. The results we obtain are presented in Table 2.

Table 2. Comparison with baselines.

	Method	V-measure	NMI	Homogeneity	Completeness
Exp 1	Baseline A1C	0.0	0.0	0.0	1.0
	Baseline MC	0.288	0.288	1.0	0.168
	Oracle Baseline	**0.539**	**0.539**	**0.722**	**0.430**
	Proposed	**0.539**	**0.539**	**0.722**	**0.430**
Exp 2	Baseline A1C	0.0	0.0	0.0	1.0
	Baseline MC	0.227	0.227	1.0	0.128
	Oracle Baseline	**0.456**	**0.456**	0.572	**0.379**
	Proposed	0.446	0.446	**0.665**	0.336

Table 3. Comparison with related work.

	Method	V-measure	NMI	Homogeneity	Completeness
Exp 1	Sridhar et al. [28]	0.098	0.103	0.076	0.138
	Proposed	**0.539**	**0.539**	**0.722**	**0.430**
Exp 2	Sridhar et al. [28]	0.128	0.133	0.100	0.176
	Proposed	**0.446**	**0.446**	**0.665**	**0.336**

7.2.1 *Results Discussion*

The baselines presented in Table 2 present the overall performance of the hierarchical clustering mechanism by evaluating the reported indices. In experiment 1, our approach reaches the optimal clustering performance, thus the reported metric values are identical to the ones of the *Oracle* baseline. In the second experiment, the proposed approach obtains performance very close to optimal clusters demonstrating a decrease of 2% for both the *v-measure* and *NMI* scores. This represents an 11% drop of the completeness score, compared to the *Oracle* baseline, causing a slight boost of 16% of the *homogeneity* score. The reported results in Table 3 demonstrate a significant elevation for all the metrics of our approach in reference to the work of Sridhar et al. [28], in both experiments.

The improvement in the reported metrics indicates the efficacy of the proposed method in creating clusters of affordances in very diverse scenarios, as well as in the presence of a temporal manifold of object interactions. In our future work we aim to evaluate our approach on a more challenging dataset with a wider range of object affordances than the ones currently considered.

7.3 Ablation Study

7.3.1 *Depth-informed Relations*

We further evaluate and provide comparisons of the proposed *DiSR* with the primary set of RCC relations for the two aforementioned experiments. We choose to compare against RCC5 which consists of the relations: 'discrete' (DR), 'partially overlapping' (PO), 'proper part' (PP, PPi), and 'equal' (EQ), since the full set of RCC-8 relations are not present in the interactions captured in CAD-120 dataset. The experimental results for both settings are presented in Table 4. The exploitation of depth information for inferring object affordances shows a considerable improvement in all metrics in comparison to using the primary RCC set. This improvement results from the more accurate spatial relationships of interactions, thus creating more accurate graphical structures for describing them.

Table 4. Experiments with and without consideration of the depth cues.

	Spatial Relations	V-measure	NMI	Homogeneity	Completeness
Exp 1	RCC5	0.337	0.337	0.359	0.317
	DiSR	**0.539**	**0.539**	**0.722**	**0.430**
Exp 2	RCC5	0.381	0.281	0.590	0.282
	DiSR	**0.446**	**0.446**	**0.665**	**0.336**

7.3.2 *Spatio-temporal Weights of Vertex Differences*

For clustering object interactions we presented the sED measure in Section 6, which is based on the normalized c_{spat} and c_{temp} parameters. These weights correspond to the impact of the spatial and temporal difference in the total graph disparity. We evaluate the contribution of the spatial and temporal relational difference to the creation of more homogeneous and complete affordance clusters by investigating all possible values for the c_{spat} and c_{temp} parameters from 0.0 to 1.0 with a step of 0.1. The results of this ablation study are illustrated in Fig. 4. For every plot, the x-axis corresponds to the height values of the produced dendrogram (*Dendrogram threshold*) from the hierarchical clustering, at which we split the dendrogram structure to create clusters with the leaf nodes. Additionally, the y-axis of the plots correlates to the suggested clustering metric (*homogeneity*, *completeness*, and *v-measure*). The label of every clustering in the figure gives the weight of the c_{spat} parameters (and the weight of the c_{temp} is thus 1 minus this value).

The results illustrate the impact the spatial and temporal relational weights have on the reported metrics. All metrics demonstrate that the best model performance is captured when we consider both spatial and temporal vertex differences. More specifically, we observe that values of c_{spat} between 0.4 and 0.6 achieve the highest *v-measure* scores. Generally, an improvement in the *homogeneity* score is evident with the increase of the c_{spat} weight, whilst acknowledging a decrease in the *completeness* score. This behavior implies that each kind of qualitative relation, *i.e.* both spatial and temporal, benefits one aspect of the clustering mechanism, albeit that both spatial and temporal information are required for more homogeneous and complete clusters of object affordances.

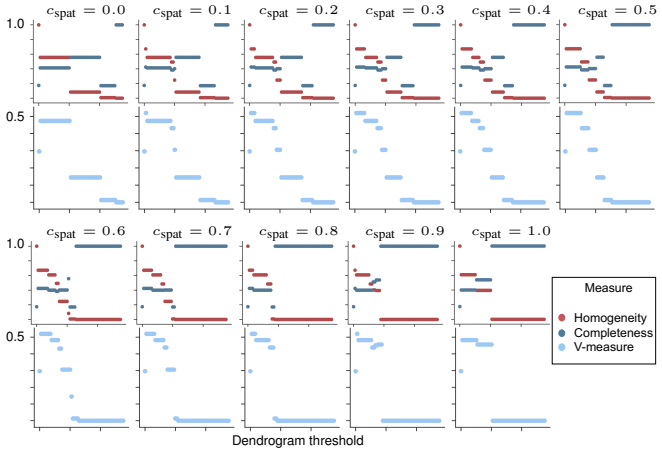

Figure 4. The *homogeneity*, *completeness*, and *v-measure* scores for the validation set with the spatial weight extending from 0.0 to 1.0.

8 LIMITATIONS

Our definitions of spatial interactions are modeling spatial states of the interactive objects, *e.g.* a cup is on the table. However, some kind of affordances are derived from interactions holding as a transition from one state to another, *e.g.* 'pourable','able to be poured' are inferred from the transition of the state Cont(bowl1, liquid) to Cont(bowl2, liquid). Our method is limited to only detect state-based relations, hence affordances as 'pourable', 'able to be poured', 'throwable', and 'able to be thrown' are not detectable in the current pipeline. Another limitation considers the ability to detect and represent visual changes of the object's state, *e.g.* deformation, cleanliness. *E.g.* 'wipeable' and 'cleanable' are affordances for which the distinction requires a different vision system and an enhanced representation of features related with the objects themselves. It is worth mentioning that, though the proposed framework is generic, it is currently limited to the set of detectable and defined relations. Moreover, an enhancement of detecting interactions of more than two objects is necessary for affordances which are inferred from the interaction of multiple objects, *e.g.* learning 'stirable' requires both a liquid and a spoon to be contained in a concave object.

9 CONCLUSION

In this work we present a novel depth-informed qualitative representation for handling occlusions and efficiently detecting relative spatial relationships between objects in the 2D image plane. We also address the problem of affordance categorization by clustering qualitative graphical structures of object interactions in an unsupervised way. Our experiments demonstrate that exploiting the proposed relations produces more accurate qualitative graphs, describing the object interactions, resulting in more homogeneous and complete clusters of affordances, and higher *v-measure* scores. Note that from the clustered interactions, it is then possible to assign object affordances to the object nodes in the $AGraphlets$ (e.g. supporter/supportee).

The enrichment of qualitative relations capturing relative motions of objects is a future direction for expanding the possible affordances our method can detect. We also aim to experiment with more diverse and complex datasets in terms of object interactions as well as enhancing the *DiSR* qualitative set with relations that also capture more complex interactions *e.g.* pouring.

REFERENCES

[1] Eren Erdal Aksoy, Alexey Abramov, Johannes Dörr, Kejun Ning, Babette Dellen, and Florentin Wörgötter, 'Learning the semantics of object–action relations by observation', *The International Journal of Robotics Research*, **30**(10), 1229–1249, (2011).

[2] Eren Erdal Aksoy, Alexey Abramov, Florentin Wörgötter, and Babette Dellen, 'Categorizing Object-Action Relations from Semantic Scene Graphs', in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pp. 398–405. IEEE, (2010).

[3] James F Allen, 'Maintaining Knowledge about Temporal Intervals', in *Readings in qualitative reasoning about physical systems*, 361–372, Elsevier, (1990).

[4] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos, 'Densepose: Dense Human Pose Estimation in the Wild', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7297–7306, (2018).

[5] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui, 'Visual Object Tracking using Adaptive Correlation Filters', in *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 2544–2550. IEEE, (2010).

[6] Anthony G Cohn, Brandon Bennett, John Gooday, and Nicholas Mark Gotts, 'Qualitative Spatial Representation and Reasoning with the Region Connection Calculus', *GeoInformatica*, **1**(3), 275–316, (1997).

[7] Thanh-Toan Do, Anh Nguyen, and Ian Reid, 'AffordanceNet: An End-to-End Deep Learning Approach for Object Affordance Detection', in *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 1–5. IEEE, (2018).

[8] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim, 'Demo2Vec: Reasoning Object Affordances from Online Videos', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2139–2147, (2018).

[9] Yiannis Gatsoulis, Muhannad Alomari, Chris Burbridge, Christian Dondrup, Paul Duckworth, Peter Lightbody, Marc Hanheide, Nick Hawes, DC Hogg, AG Cohn, et al., 'QSRlib: A Software Library for Online Acquisition of Qualitative Spatial Relations from Video', *In Workshop on Qualitative Reasoning (QR16), at IJCAI*, (2016).

[10] James J Gibson, 'The Theory of Affordances', *Hilldale, USA*, **1**(2), (1977).

[11] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He, 'Detecting and Recognizing Human-Object Interactions', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8359–8367, (2018).

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, 'Mask R-CNN', in *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 2980–2988. IEEE, (2017).

[13] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena, 'Learning Human Activities and Object Affordances from RGB-D videos', *The International Journal of Robotics Research*, **32**(8), 951–970, (2013).

[14] Michael Leyton, 'A Process-Grammar for Shape', *Artificial Intelligence*, **34**(2), 213–247, (1988).

[15] Wei Liang, Yibiao Zhao, Yixin Zhu, and Song-Chun Zhu, 'What Is Where: Inferring Containment Relations from Videos', in *IJCAI*, pp. 3418–3424, (2016).

[16] Wei Liang, Yixin Zhu, and Song-Chun Zhu, 'Tracking Occluded Objects and Recovering Incomplete Trajectories by Reasoning about Containment Relations and Human Actions', in *Thirty-Second AAAI Conference on Artificial Intelligence*, (2018).

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, 'Microsoft COCO: Common Objects in Context', in *European conference on computer vision*, pp. 740–755. Springer, (2014).

[18] Bogdan Moldovan and Luc De Raedt, 'Occluded Object Search by Relational Affordances', in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 169–174. IEEE, (2014).

[19] Luis Montesano and Manuel Lopes, 'Learning Grasping Affordances from Local Visual Descriptors', in *2009 IEEE 8th international conference on development and learning*, pp. 1–6. IEEE, (2009).

[20] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos, 'Affordance Detection of Tool Parts from Geometric Features', in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1374–1381. IEEE, (2015).

[21] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis, 'Detecting Object Affordances with Convolutional Neural Networks', in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2765–2770. IEEE, (2016).

[22] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis, 'Object-Based Affordances Detection with Convolutional Neural Networks and Dense Conditional Random Fields', in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5908–5915. IEEE, (2017).

[23] Alessandro Pieropan, Carl Henrik Ek, and Hedvig Kjellström, 'Functional Object Descriptors for Human Activity Modeling', in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pp. 1282–1289. IEEE, (2013).

[24] Alessandro Pieropan, Carl Henrik Ek, and Hedvig Kjellström, 'Recognizing Object Affordances in Terms of Spatio-Temporal Object-Object Relationships', in *International Conference on Humanoid Robots, November 18-20th 2014, Madrid, Spain*, pp. 52–58. IEEE conference proceedings, (2014).

[25] Siyuan Qi, Yixin Zhu, Siyuan Huang, Chenfanfu Jiang, and Song-Chun Zhu, 'Human-Centric Indoor Scene Synthesis Using Stochastic Grammar', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5899–5908, (2018).

[26] David A Randell, Zhan Cui, and Anthony G Cohn, 'A Spatial Logic based on Regions and Connection', *KR*, **92**, 165–176, (1992).

[27] Johann Sawatzky, Abhilash Srikantha, and Juergen Gall, 'Weakly Supervised Affordance Detection', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2795–2804, (2017).

[28] Muralikrishna Sridhar, Anthony G Cohn, and David C Hogg, 'Learning Functional Object Categories from a Relational Spatio-Temporal Representation', in *ECAI 2008: 18th European Conference on Artificial Intelligence (Frontiers in Artificial Intelligence and Applications)*, pp. 606–610. IOS Press, (2008).

[29] Muralikrishna Sridhar, Anthony G Cohn, and David C Hogg, 'Relational Graph Mining for Learning Events from Video', in *Proceedings of the 2010 conference on STAIRS 2010: Proceedings of the Fifth Starting AI Researchers Symposium*, pp. 315–327, (2010).

[30] Alexia Toumpa and A Cohn, 'Relational Graph Representation Learning for Predicting Object Affordances', in *NeurIPS Workshop on Graph Representation Learning*, (2019).

[31] Matthew W Turek, Anthony Hoogs, and Roderic Collins, 'Unsupervised Learning of Functional Categories in Video Scenes', in *European Conference on Computer Vision*, pp. 664–677. Springer, (2010).

[32] Bangpeng Yao, Jiayuan Ma, and Li Fei-Fei, 'Discovering Object Functionality', in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2512–2519, (2013).

[33] Yibiao Zhao and Song-Chun Zhu, 'Scene Parsing by Integrating Function, Geometry and Appearance Models', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3119–3126, (2013).