UNIVERSITY OF LEEDS

Categorization of Affordances and Prediction of Future Object Interactions using Qualitative Spatial Relations

Alexia Toumpa

Submitted in accordance with the requirements for the degree of Doctor of Philosophy (PhD)

The University of Leeds

Faculty of Engineering and Physical Sciences

School of Computing

July 2023

Intellectual Property

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given where reference has been made to the work of others.

Some parts of the work presented in, Chapter 3, Chapter 4, Chapter 5, and Chapter 6 of this thesis have been published in the following articles. The publications are primarily the work of the candidate.

Alexia Toumpa and Anthony G Cohn (2023a). "Future Qualitative Activity Graph Prediction". In: *37th AAAI Conference on Artificial Intelligence, 3rd Workshop on Graphs and more Complex Structures for Learning and Reasoning (GCLR)*

Alexia Toumpa and Anthony G Cohn (2023b). "Object-agnostic Affordance Categorization via Unsupervised Learning of Graph Embeddings". In: *Journal of Artificial Intelligence Research* 77, pp. 1–38

Alexia Toumpa and Anthony G Cohn (2020). "Depth-informed Qualitative Spatial Representations for Object Affordance Prediction". In: *33rd International Workshop on Qualitative Reasoning*

Alexia Toumpa and Anthony G Cohn (2019). "Relational Graph Representation Learning for Predicting Object Affordances". In: *NeurIPS Workshop on Graph Representation Learning*

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

© 2023 The University of Leeds, Alexia Toumpa

Acknowledgements

I gratefully acknowledge the financial support provided by the University of Leeds and the Engineering and Physical Sciences Research Council (EPSRC). I would especially like to thank my advisor Professor Anthony Cohn for his guidance and support, as well as Professor David Hogg for his valuable input. Also, I would like to thank my viva examiners Dr. Brandon Bennett and Professor Artur d' Avila Garcez for our insightful discussion, as well as their comments and feedback.

In addition, I would like to thank my friends and colleagues in the School of Computing and the Robotics Laboratory for a cherished time spent together in the lab and in social settings, as well as Panagiotis Magkafas for his technical assistance in re-implementing the work of Aksoy, Abramov, Wörgötter, et al. 2010.

Also, I would like to extend my sincere thanks to Alexandros, who has been my great companion, for his endless patience, encouragement, motivation, and for making those years so enjoyable. Lastly, I would like to express my profound gratitude to my father Yiannis, my mother Aleka, and my sister Dimitra for their unwavering support throughout this journey.

This accomplishment would not have been possible without them. Thank you.

Abstract

The application of deep neural networks on robotic platforms has successfully advanced robot perception in tasks related to human-robot collaboration scenarios. Tasks such as scene understanding, object categorization, affordance detection, interaction anticipation, are facilitated by the acquisition of knowledge about the object interactions taking place in the scene.

The contributions of this thesis are two-fold:

1. it shows how representations of object interactions learned in an unsupervised way can be used to predict categories of objects depending on the affordances;

2. it shows how future frame-independent interaction can be learned in a self-supervised way by exploiting high-level graph representations of the object interactions.

The aim of this research is to create representations and perform predictions of interactions which abstract from the image space and attain generalization across various scenes and objects. Interactions can be static, *e.g.* holding a bottle, as well as dynamic, *e.g.* playing with a ball, where the temporal aspect of the sequence of several static interactions is of importance to make the dynamic interaction distinguishable. Moreover, occlusion of objects in the 2D domain should be handled to avoid false positive interaction detections. Thus, RGB-D[1] video data is exploited for these 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[2] and class-agnostic objects[3]. In order

[1]RGB-D stands for Red Green Blue - Depth, and provides RGB as well as the corresponding depth information of an image.

[2]Specifically in Machine Learning, *open-set* refers to the task of identifying entities, *e.g.* interactions, objects, from a predefined set of classes, as well as distinguishing those that do not fall into any of the defined classes.

[3]In the context of object detection, class-agnostic objects refer to object proposals without their association to an object class. Hence, they carry only the information of the detection area/region of a potential object in

to abstract from the continuous representation of spatio-temporal interactions in video data, a novel set of high-level qualitative depth-informed spatial relations is presented. Learning similarities via an unsupervised method exploiting graph representations of object interactions induces a hierarchy of clusters of objects with similar affordances. The proposed method handles object occlusions by capturing effectively possible interactions and without imposing any object or scene constraints.

Moreover, interaction and action anticipation remains a challenging problem, especially considering the generalizability constraints of trained models from visual data or exploiting visual video embeddings. State of the art methods allow predictions approximately up to three seconds of time in the future. Hence, most everyday-life activities, which consist of actions of more than five seconds in duration, are not predictable. This thesis presents a novel approach for solving the task of interaction anticipation between objects in a video scene by utilizing high-level qualitative frame-number-independent spatial graphs to represent object interactions. A deep recurrent neural network learns in a self-supervised way to predict graph structures of future object interactions, whilst being decoupled from the visual information, the underlying activity, and the duration of each interaction taking place.

Finally, the proposed methods are evaluated on RGB-D video datasets capturing everyday-life activities of human agents, and are compared against closely-related and state-of-the-art methods.

the scene on the visual information.

Contents

List of Figures

List of Tables

List of Symbols

b	batch size
b_α	bias specifically for α, where $\alpha \in \{i, f, c, o\}$
C	a group of clusters
C_C	cohesion of a partition C
C_t	memory cell state at time t of an LSTM cell
$\bar{c_k}$	the centroid of a cluster c_k
$cos(\cdot)$	cosine function
D	a dataset
D_C	density of a partition C
d_cmin	minimum value from a set of depth values indicating a concave curve
d_cmax	maximum value from a set of depth values indicating a concave curve
$\text{dist}_{\text{depth}}$	increasing-ordered-by-value depth information
$dmax$	maximum value from a set of depth values
$dmin$	minimum value from a set of depth values
E	edges of a graph G
e	episode of interaction
F	number of predicted relations, used for the dimensionality of the output of a neural network
f_t	forget gate at time t of an LSTM cell
G	a graph
H_t	hidden state of a neural network layer at time t
i_t	input gate at time t of an LSTM cell
$\mathscr{J}$	Jaccard index similarity measure function
K	kernel size
$\mathfrak{L}$	loss/cost function

M	number of samples in a batch of data
mask(x)	mask area of an entity x
n	the number of objects in a dataset
n_α	the number of objects from a set α
O	number of detectable objects in a scene, using an off-the-shelf object detector
O_C	overlap of a partition C
o_t	output gate at time t of an LSTM cell
P	number of input relations, used for the dimensionality of input data in a neural network
$\mathbb{R}$	set of real numbers
S_C	separation of a partition C
sd	number of sections the depth information of an object can be partitioned in
sd_{max}	the number of sections in the ordered-by-value depth information with the highest values
T_t	input tensor, at time t, of a neural network
$thresh_{convex}$	threshold depth value that distinguishes “convex” from “concave” objects
V	vertices of a graph G
V_C	variance of a partition C
V_l	the vertices of the l layer of an Activity Graph, where $l = \{ent, spat, temp\}$
$V_l^{'}$	the vertices of the l layer of an Activity Graphlet, where $l = \{ent, spat, temp\}$
W_α	weights of a neural network specific for α, where $\alpha \in \{x, h, o\}$
X	the objects of a dataset, where $X = \{x_1, \ldots, x_i, \ldots, x_n\}$ and $1 < i < n$
X_t	input at time t of a neural network
Δ	the diameter of a cluster
δ	set distance function
$\sigma(\cdot)$	the sigmoid function
$\lvert \cdot \rvert$	the number of elements in a set

List of Acronyms

AG	Activity Graph
AGraphlet	Activity Graphlet
C	Connected (RCC relation)
CAD	CAD-120 Dataset
CarDir	Cardinal Direction
Cont, Conti	Containing, Containing inverse (DiSR relation)
ConvLSTM	Convolutional Long Short-Term Memory
CPM	Convolutional Pose Machine
DC	Disconnected (RCC relation)
DiSR	Depth-informed Spatial Relations
DR	Discrete (RCC relation)
d, di	during, during inverse (Allen's relation)
e.g.	exempli gratia
EC	Externally Connected (RCC relation)
EQ	Equal (RCC relation)
etc	Et cetera
f, fi	finishes, finishes inverse (Allen's relation)
FC	Fully Connected
i.e.	id est
LOAD	Leeds Object Affordance Dataset
LSTM	Long Short-Term Memory
m, mi	meets, meets inverse (Allen's relation)
M+	protrusion area
m-	indentation area

MoS	Moving or Stationary
NAdj	Not Adjacent (DiSR relation)
NTPP, NTPPi	Non-Tangential Proper Part, Non-Tangential Proper Part inverse (RCC relation)
O	Overlaps (RCC relation)
o, oi	overlaps, overlaps inverse (Allen's relation)
P, Pi	Part, Part inverse (RCC relation)
PO	Partially Overlapping (RCC relation)
PP, PPi	Proper Part, Proper Part inverse (RCC relation)
QSR	Qualitative Spatial Relations
QTC	Qualitative Trajectory Calculus
RCC	Region Connection Calculus
RCC2	Region Connection Calculus with two relations
RCC5	Region Connection Calculus with five relations
RCC8	Region Connection Calculus with eight relations
RGB	Red Green Blue
RGB-D	Red Green Blue Depth
s, si	starts, starts inverse (Allen's relation)
sED	set Edit Distance
Sup, Supi	Supporting, Supporting inverse (DiSR relation)
T	Touching (DiSR relation)
TPP, TPPi	Tangential Proper Part, Tangential Proper Part inverse (RCC relation)
WnP	Watch-n-Patch Dataset
<, >	before, before inverse (Allen's relation)
=	equals (Allen's relation)

Chapter 1

Introduction

1.1 Motivation

Humans interact with objects in a real-world environment depending on the purpose of the underlying activity taking place and the availability of the objects in the scene. Humans can be quite inventive when it comes to performing an activity with limited object availability. *E.g.* a book can be used as a tray when no flat surface is available, a suitcase can be used as a cover when no other covering object is obtainable. Thus, objects may be used differently from the purpose of their creation, *i.e.* the book is created for writing/reading, a suitcase is created for containing.

In an everyday-life scenario humans interact in multiple ways with different objects. Every object can have various affordances, according to its shape and the scene it is present in. The *affordance* of an object[1] essentially describes the way it can be used by an agent. *E.g.* the microwave can be used as a container and as a supporter, a table can be used as a supporter, as a coverer, as well as a surface to sit on. However, the *functionality* is the characteristic of an object that describes its main usage considering the purpose of its creation. *E.g.* a chair is used for sitting, a microwave is used for containing, a table is used for supporting.

Humans tend to describe these concepts by using qualitative relations between artifacts, for instance, "the cat is under the table" instead of giving the exact coordinates of the world and the object. From an engineering perspective, an interesting approach to formulate the human understanding of the affordances and functionality of the objects in a scene is by using qualitative

[1] A formal definition of the term affordance is given in Sec. 5.1

spatio-temporal relations. These qualitative relations, provide generalization across different scenarios. However, they need to be able to capture correctly the object interactions present in a real-world environment, thus occlusion needs to be addressed.

Furthermore, when it comes to human-robot collaboration scenarios, understanding and predicting object affordances is a crucial task. Previous research has limited the number and kind of objects available in the scene, and understanding the object affordances has been highly correlated with the underlying activity and the utilized object. This thesis introduces a new approach which relies on the object interactions to extract object affordances and performs generalizable predictions for open-set object detections[2], scenes, and activities.

In a human-robot collaboration scenario, apart from acquiring information about how the objects can be employed in a scene, another key task is the ability to predict future actions/interactions. *E.g.* in an assembly scenario, the robot's perception of the action sequence and the upcoming actions is crucial for interacting with a human.

The prediction of future actions/interactions is correlated with the human intention. The latter is an aspect of perception which cannot be extracted from low-level features, *i.e.* pixel values, thus a high-level representation of states needs to be formulated.

For human beings, it seems to be an easy task to understand an action, its purpose and the reason some objects might be used. However, current AI-based approaches cannot infer the high-level knowledge behind the scope of an action as a human can. Understanding the purpose of an action is highly correlated with predicting the future action after it has been performed.

This thesis presents a novel approach for interaction prediction without acquiring any knowledge about the intention of the human in the scene, but by defining and predicting the purpose of an action performed by interpreting the world states, *i.e.* sequences of spatio-temporal interactions, with high-level representations.

There is a wide literature addressing the problem of interaction anticipation in terms of predicting future visual representations of the scene. However, these approaches lack the ability to predict interactions in really-long term activities. The proposed method addresses this limitation by acquiring high-level information of interactions in the domain of space and time.

[2]Open-set object detections, in the context of object recognition and classification, considers object detections of known as well as unknown object categories, *i.e.* classes.

Among the many challenging tasks involved in human-robot collaboration scenarios, this thesis addresses the problem of categorizing previously unseen objects based on the way they can be used in any scenario and activity taking place, and anticipating future interactions without posing any number of frame, scene, object and activity restrictions.

1.2 Thesis Aim

This thesis presents a novel approach for learning to categorise object affordances in an unsupervised way by solely relying on qualitative information describing the objects' interactions. To provide better interaction descriptions and representations, a novel set of qualitative spatial relations is introduced. This relational set is able to represent more complex spatial relations than existing ones, by incorporating information about the objects' shape to infer relations. Moreover, this thesis introduces a new research direction for predicting future interactions by exploiting qualitative spatial graph representations of interactions.

A new video dataset capturing human-object interactions in everyday-life activities has been made publicly available to the community, and the newly presented set of relations has been integrated into an existing open source library. Finally, this thesis aims to inspire new research directions in learning interactions and predicting interactions through graph data extracted from videos, by utilizing more complex information, *e.g.* relations, rather than pixel values. It also aims to motivate research on the semantics of interactions and understanding the purpose of human actions.

1.3 Research Contributions

The main research contributions of this thesis are:

1. a novel depth-informed mereotopological set of relations, that captures complex spatial relationships, *i.e.* "supporting", "containing", and enables the distinction between occlusions and interactions.;
2. a newly introduced RGB-D video dataset that captures everyday-life activities with human-object interactions, expanding the domain of object affordances;
3. an unsupervised learning method on high-level qualitative graph embeddings for creating categories of object affordances, without posing any scene or object restrictions. The

categories of affordances form a hierarchy of groups of similar object affordances based on the observed object interactions;

4. a self-supervised method based on a Convolutional LSTM model for predicting number-of-frame-independent future qualitative graphs of spatial object interactions, whilst being object and scene independent.

1.4 Published Work

The work presented in this thesis is supported by the following publications:

Alexia Toumpa and Anthony G Cohn (2023a). "Future Qualitative Activity Graph Prediction". In: *37th AAAI Conference on Artificial Intelligence, 3rd Workshop on Graphs and more Complex Structures for Learning and Reasoning (GCLR)*

Alexia Toumpa and Anthony G Cohn (2023b). "Object-agnostic Affordance Categorization via Unsupervised Learning of Graph Embeddings". In: *Journal of Artificial Intelligence Research* 77, pp. 1–38

Alexia Toumpa and Anthony G Cohn (2020). "Depth-informed Qualitative Spatial Representations for Object Affordance Prediction". In: *33rd International Workshop on Qualitative Reasoning*

Alexia Toumpa and Anthony G Cohn (2019). "Relational Graph Representation Learning for Predicting Object Affordances". In: *NeurIPS Workshop on Graph Representation Learning*

1.5 Organization

The rest of this thesis is organized as follows. Firstly, Chapter 2 presents all the background knowledge employed for representing interactions and exploiting various metrics for evaluating the performance of classification. In Chapter 3 a novel set of depth-informed spatial relations is introduced, which uses the objects' convexity-type and their depth information to detect when and which relation holds at every time interval. This set of spatial relations is employed in Chapter 5 for capturing object-object interactions and learning in an unsupervised way categories

of object affordances. For the evaluation of the proposed method a newly introduced RGB-D video dataset is used, presented in Chapter 4. Moreover, this thesis proposes a new approach for predicting future interactions between entities in a video scene by exploiting qualitative spatio-temporal graphs. Chapter 6 presents a new approach which employs a Convolutional LSTM model to perform predictions of future qualitative spatio-temporal graphical structures.

Finally, Chapter 7 makes some concluding remarks drawn from this thesis along with a discussion of some future research directions.

Chapter 2

Background

2.1 Introduction

This chapter sets out the basis for the rest of this thesis, by presenting all the background knowledge one needs to have before continuing to the next chapters. It introduces some of the most important qualitative spatio-temporal relations and presents how graphical structures are created using these qualitative relations to describe interactions present in video data. In Chapter 3, Chapter 5 and Chapter 6 these qualitative relations and graph structures are exploited to represent object interactions.

Moreover, this chapter provides an overview of the evaluation metrics for the task of classification. Depending on the ground truth available metrics are split into three categories: *internal*, *external*, and *relative* validation indices. In Chapter 5 the output of an unsupervised learning algorithm is evaluated by considering the *external* validation indices presented here.

This chapter is organized as follows. Qualitative spatio-temporal relations are introduced in Section 2.2, focusing on the relations that will be exploited throughout this thesis. These qualitative relations are further employed to create *Activity Graphs*, presented in Section 2.3.

Furthermore, Section 2.4 summarizes some of the most important *internal*, *external*, and *relative* validation indices for evaluating classification outputs.

2.2 Qualitative Relations

Qualitative Spatio-temporal Relations (QSRs) describe spatial and temporal relationships between entities in space and time. Spatial relational sets are Joint Exhaustive and Pairwise Disjoint (JEPD) sets, *i.e.* for any kind of spatial entities there is one and only one spatial relation from the set of relations that can hold. Moreover, transitions between QSR relations exist. A *conceptual neighborhood* (Freksa 1992; Gooday and A. Cohn 1994) is a directed graph of transitions between the relations. Neighboring pairs of relations in such a graph are called *conceptual neighbors*. Furthermore, qualitative spatial relations are divided into mereotopology, direction, distance, shape, and moving objects.

Mereotopological relations describe the parthood and topology of entities, and are the most well-studied QSRs. The two best-known approaches for representing and reasoning with topological relations are the Region Connection Calculus (RCC) (Anthony G Cohn et al. 1997; Gerevini and Renz 2002) and the n-intersections (Egenhofer and Franzosa 1991; Egenhofer, Clementini, et al. 1994; Egenhofer and Franzosa 1995; Egenhofer 2005; Egenhofer and Herring 1990; Egenhofer, Mark, et al. 1994; Egenhofer and Sharma 1993; Egenhofer and Vasardani 2007).

Direction relations describes where an entity is located in space relative to another entity. These relations are more constrained than topological relations. Cone-shaped direction relations (Isli et al. 2001), projection-based direction relations (Isli et al. 2001), the Oriented-Point Algebra (OPRA) (Moratz 2006), the Ternary-Point Configuration Calculus (TPCC) (Moratz and Ragni 2008), and the Cardinal Direction Calculus (CDC) (Skiadopoulos and Koubarakis 2004; Skiadopoulos and Koubarakis 2005) are some of the calculi that describe the relative direction information between two entities.

Distance spatial relations can be *absolute*, *i.e.* the actual measured distance between two entities, or *relative*, *i.e.* comparison of distance measure using a third entity. Distance relations are not very informative on their own, hence they are used along with the direction relations.

Shape relations aim to describe in a qualitative way the shape of entities. They can be region-based methods (Anthony G Cohn 1995), boundary-based methods, *i.e.* Process-Grammar (Leyton 1988), or string representation methods, *e.g.* qualitative curvature types (Galton and Meathrel 1999).

Moving objects relations focus on the relationship between trajectories of moving point entities.

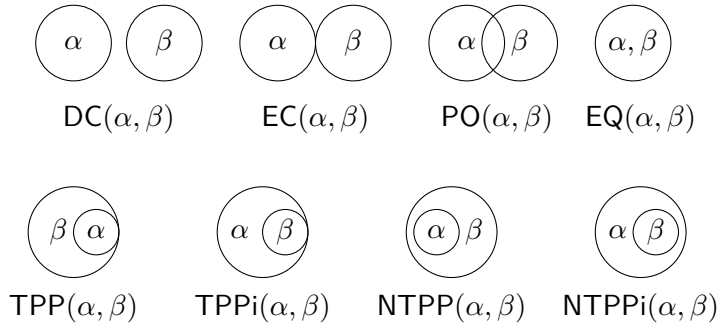
The family of Qualitative Trajectory Calculi (QTC) (Van de Weghe, Anthony G Cohn, De Tre, et al. 2006; Delafontaine et al. 2011) provide trajectory comparison by considering relative motions.

This section focuses on the topological relations RCC, the moving objects relations QTC and relations from a temporal algebra. The selection of these relational sets was based on their ability to describe and represent sequences of interactions between entities.

2.2.1 Region Connection Calculus

The *Region Connection Calculus* (RCC) (Anthony G Cohn et al. 1997; Gerevini and Renz 2002) is one of the best known approaches for representing and reasoning of topological relations between entities in space. The $\mathsf{C}(\alpha, \beta)$ relation is used to define the RCC set of relations (Table 2.1). This relation holds between two non-empty regions[1] (α and β) and is interpreted as the connection of these entities if and only if their topological closures share a common point.

The RCC8 set contains the relations: disconnected (DC), externally connected (EC), partially overlapping (PO), equal (EQ), tangential proper part (TPP), tangential proper part inverse (TPPi), non-tangential proper part (NTPP), and non-tangential proper part inverse (NTPPi), which are visualized below:



where α and β are two spatial entities.

The conceptual neighborhood of the RCC8 relations is illustrated in Figure 2.1.

RCC2 is the primitive set of RCC relations, containing the relations: C and DC, and RCC5 set is a coarser set of the RCC8 comprising the relations: DR, PO, PP, PPi, and EQ, and their mapping to the RCC8 set is visualized below:

[1] Non-empty regions imply that they correspond to solid objects, rather than point objects.

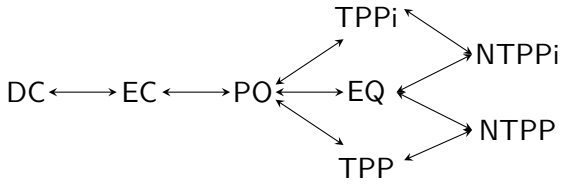
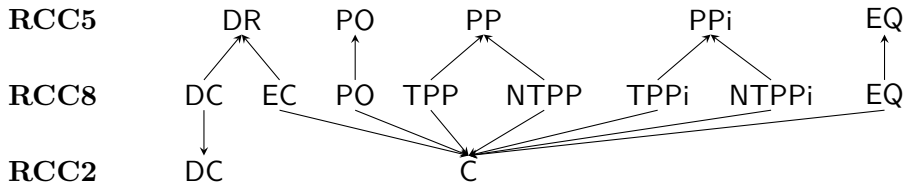


Figure 2.1: Conceptual neighborhood of RCC8.

RCC	**Definition**	**Description**
$\mathsf{DC}(\alpha,\beta)$	$\neg\mathsf{C}(\alpha,\beta)$	α is disconnected from β
$\mathsf{P}(\alpha,\beta)$	$\forall\gamma(\mathsf{C}(\gamma,\alpha)\rightarrow\mathsf{C}(\gamma,\beta))$	α is a part of β
$\mathsf{Pi}(\alpha,\beta)$	$\mathsf{P}(\beta,\alpha)$	β is a part of α
$\mathsf{PP}(\alpha,\beta)$	$\mathsf{P}(\alpha,\beta)\wedge\neg\mathsf{P}(\beta,\alpha)$	α is a proper part of β
$\mathsf{PPi}(\alpha,\beta)$	$\mathsf{PPi}(\beta,\alpha)$	β is a proper part of α
$\mathsf{EQ}(\alpha,\beta)$	$\mathsf{P}(\alpha,\beta)\wedge\mathsf{P}(\beta,\alpha)$	α equals β
$\mathsf{O}(\alpha,\beta)$	$\exists\gamma(\mathsf{P}(\gamma,\alpha)\wedge\mathsf{P}(\gamma,\beta))$	α overlaps β
$\mathsf{PO}(\alpha,\beta)$	$\mathsf{O}(\alpha,\beta)\wedge\neg\mathsf{P}(\alpha,\beta))\wedge\neg\mathsf{P}(\beta,\alpha)$	α partially overlaps β
$\mathsf{DR}(\alpha,\beta)$	$\neg\mathsf{O}(\alpha,\beta)$	α is discrete from β
$\mathsf{EC}(\alpha,\beta)$	$\mathsf{C}(\alpha,\beta)\wedge\neg\mathsf{O}(\alpha,\beta)$	α is externally connected with β
$\mathsf{TPP}(\alpha,\beta)$	$\mathsf{PP}(\alpha,\beta)\wedge\exists\gamma(\mathsf{EC}(\gamma,\alpha)\wedge\mathsf{EC}(\gamma,\beta))$	α is a tangential proper part of β
$\mathsf{TPPi}(\alpha,\beta)$	$\mathsf{TPP}(\beta,\alpha)$	β is a tangential proper part of α
$\mathsf{NTPP}(\alpha,\beta)$	$\mathsf{PP}(\alpha,\beta)\wedge\neg\exists\gamma(\mathsf{EC}(\gamma,\alpha)\wedge\mathsf{EC}(\gamma,\beta))$	α is a non-tangential proper part of β
$\mathsf{NTPPi}(\alpha,\beta)$	$\mathsf{NTPP}(\beta,\alpha)$	β is a non-tangential proper part of α

Table 2.1: RCC definitions.



2.2.2 Qualitative Trajectory Calculus

The basic *Qualitative Trajectory Calculus* (QTC) (Van de Weghe, Anthony G Cohn, De Tre, et al. 2006; Delafontaine et al. 2011) compares motion trajectories of relative point-entities. The relations comprising the QTC set are: 0, +, and -, denoting three motion states: stationary, moving away, and moving towards. Considering a pair-wise set of entities, the basic QTC relations are shown in Figure 2.2, where a black-filled and non-filled circle denotes a stationary

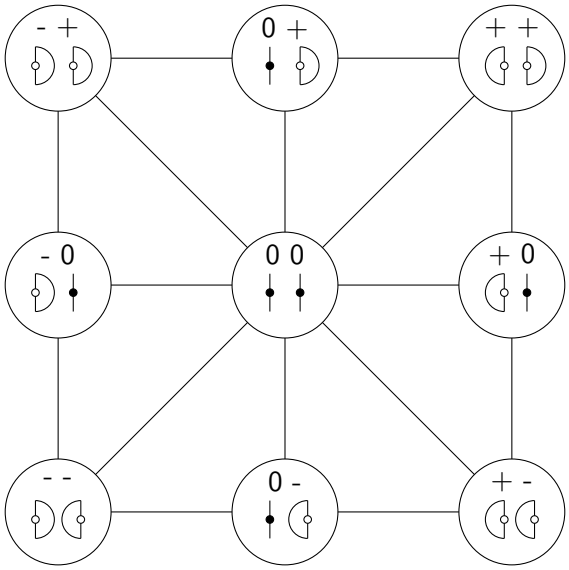


Figure 2.2: QTC relations (Van de Weghe, Anthony G Cohn, and De Maeyer 2004).

and moving entity, respectively. The direction of the arc indicates the direction of motion of a moving entity, and the line notation implies that there is no motion towards either direction for the corresponding object. Moreover, the links between the different QTC states illustrate the conceptual neighborhood.

2.2.3 Allen's Interval Algebra

The *Allen's interval algebra* (Allen 1990) is a set of temporal relations that describe in a qualitative way the temporal interactions of event intervals. Table 2.2 presents the relations that comprise the *Allen's interval algebra* along with their descriptions.

2.3 Activity Graphs

Graph structures are able to capture high level information of relations or even dynamic relational changes, *e.g.* a spatial relation between two entities or their spatio-temporal relational change, while interacting.

An *Activity Graph* (*AG*) (Sridhar et al. 2010a), is a graph representation which captures spatio-temporal information of the interactions between entities present in a sequence of interactions. Let $G = (V, E)$ be an *AG*, where the vertices V are partitioned into 3 layers: the *entity* layer, the *spatial* layer, and the *temporal* layer. The edges E exists only between adjacent layers to represent pair-wise entity interactions. Each layer of vertices comprises a single type of node:

Relation	Description	Pictorial example
$\alpha < \beta$, $\beta > \alpha$	event α happens before event β	$\alpha\ \alpha\ \alpha \qquad \beta\ \beta$
α m β, β mi α	event α meets event β	$\alpha\ \alpha\ \alpha\ \beta\ \beta$
α o β, β oi α	event α overlaps with event β	$\alpha\ \alpha\ \alpha$ $\beta\ \beta\ \beta\ \beta\ \beta$
α s β, β si α	event α starts with event β	$\alpha\ \alpha\ \alpha$ $\beta\ \beta\ \beta\ \beta\ \beta$
α d β, β di α	event α happens during event β	$\alpha\ \alpha\ \alpha$ $\beta\ \beta\ \beta\ \beta\ \beta\ \beta$
α f β, β fi α	event α finishes with event β	$\alpha\ \alpha\ \alpha$ $\beta\ \beta\ \beta\ \beta\ \beta$
$\alpha = \beta$	event α happens at the same time with event β	$\alpha\ \alpha\ \alpha$ $\beta\ \beta\ \beta$

Table 2.2: Allen's interval algebra.

1. the *entity* layer, contains the set of vertices of the entities which interact (V_{ent}),
2. the *spatial* layer, consists of vertices with the spatial relations (V_{spat}) which describe the spatial relations of the entities in V_{ent}, and
3. the *temporal* layer, specifies the temporal relations between the *spatial* layer nodes (V_{temp}).

For the *spatial* layer, qualitative spatial relations can be exploited (Sec. 2.2) and for the *temporal* layer the *Allen's Interval Algebra* (Allen 1990) is most commonly used.

The sequence of spatial relations obtained in every pair-wise interaction is extracted from the presence of episodes. An *episode* (e) represents a period of time throughout which a spatial relation between the pair of entities occurs, whilst before and after the defined episode a different spatial relation holds. The *temporal* layer encodes the temporal relationships between the episodes over which particular spatial relationships hold.

Considering the RCC8 relations for describing spatial relationships between interacting entities, Figure 2.3 illustrates the detected episodes of an example sequence of interactions between three spatial solid entities α, β, and γ. Figure 2.4 presents the RCC8 relations holding between α, β, and γ, for the episodes of Figure 2.3.

An example of three spatial entities α, β, and γ interacting, is presented in Figure 2.5. The *AG* illustrated considers the RCC8 relations to describe the entities' topological relations, and

Allen's temporal algebra to describe the temporal sequence of spatial interactions, following the detection of episodes. The direction of the edges in the graph, captures the nodes' correlation. *I.e.* the starting nodes of incoming edges of the V_{spat} and V_{temp} nodes represent the first argument of the relation of the end node of such edges, whereas the end nodes of outgoing edges of the V_{spat} and V_{temp} nodes represent the second argument of the relation of the starting node of such edges.

2.4 Evaluation Metrics

When implementing unsupervised learning approaches, having a labelled dataset to perform various comparisons to evaluate the performance of the classification structure is not always feasible and it is considered to be a major challenge in research. As it is commonly known, an unsupervised classification method that separates the input data into groups is called clustering. The objective of clustering algorithms is to define partitions where the data within the clusters display similarities. Therefore, the question of how to estimate a clustering structure as significant, arises.

Depending on the information available for the validation process, cluster validation techniques are classified into three groups (Arbelaitz et al. 2013):

1. *internal validation*, if there are no reference partitions;
2. *external validation*, if there exists a reference partition, *i.e.* ground truth;
3. *relative validation*, if synthesised partitions can be generated by running the clustering algorithm on the original dataset or on a subset of the dataset.

When a reference partition is provided, *i.e. external validation*, then the evaluation of a clustering structure can be done by comparing it with the proposed partitioning structure from the clustering algorithm. However, when a reference partitioning structure is not known, *i.e. internal validation*, then there are two ways to perform a cluster validity check: either by focusing on the partitioned data and measuring the *compactness* (or *cohesion*) and *separation* of the clusters, or by performing *Stability-based Validation*, which relies on the stability of the clustering algorithm over different samples of the input data.

Most cluster validation indices are defined by employing the *compactness* and *separability* as

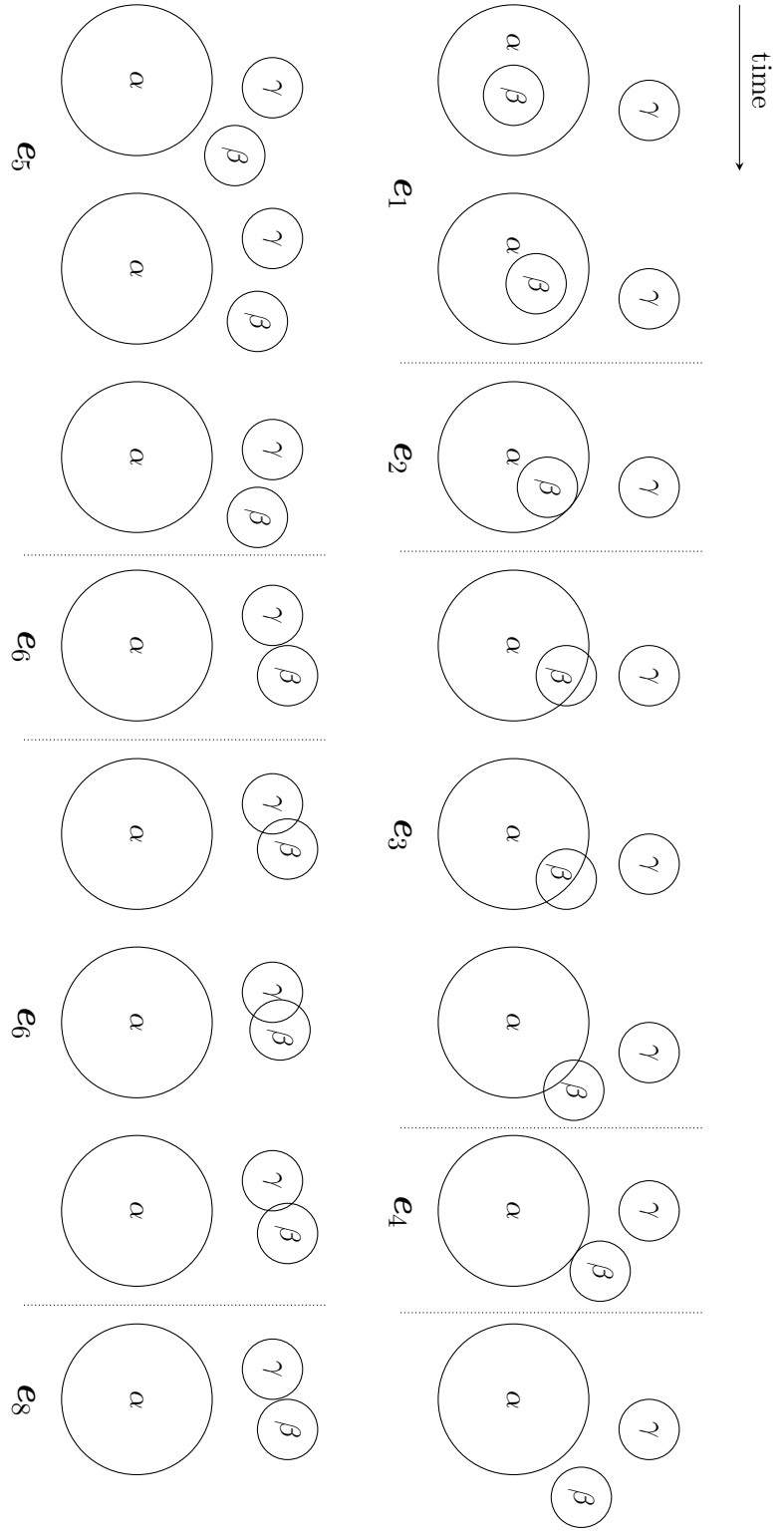


Figure 2.3: Example sequence of interactions between objects α, β, and γ, and their detected episodes (e).

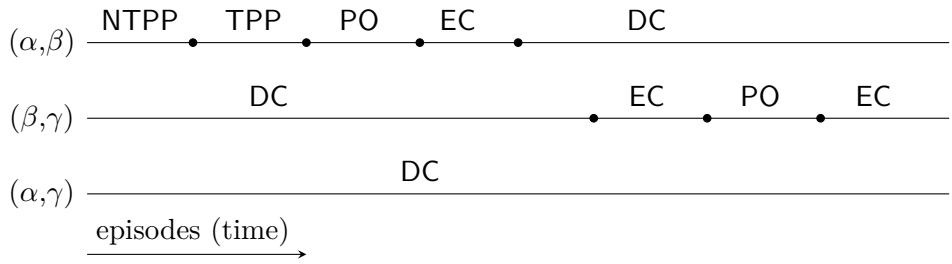


Figure 2.4: RCC8 relations between α, β, and γ, based on their interactions from Figure 2.3.

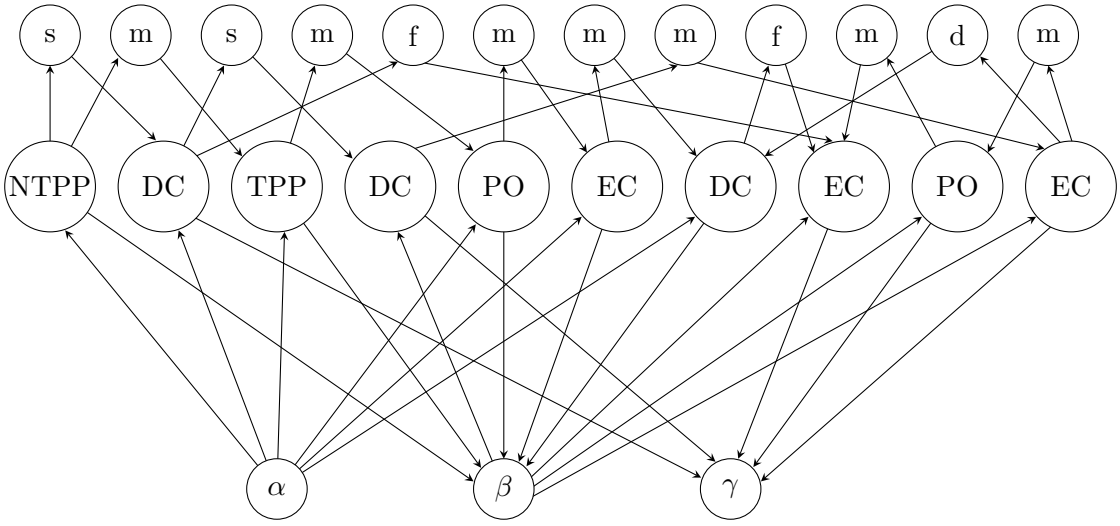


Figure 2.5: Activity graph for the α, β, and γ entities. The construction of the activity graph is based on the episode and relation detection between the interacting entities (Fig 2.3 and Fig. 2.4).

evaluation criteria (Berry and Linoff 2004). *Compactness* (or *cohesion*) measures how close the elements of a cluster are, *i.e.* variance. A low value of variance indicates closeness of the data points. *Separability* is a measure of how distinct clusters of a partition are.

The rest of this section focuses on the introduction of the external cluster validation indices, since they are employed in the following chapters. An introduction and extensive analysis of the internal cluster validation approach from the literature can be found in Appendix Section A.1.

2.4.1 Notations

Let it be a dataset D which consists of n objects: $X = \{x_1, ..., x_i, ..., x_n\}$, $1 < i < n$. A partition C of X defines a set of disjoint clusters which partition X into K groups: $C = \{c_1, ..., c_k, ..., c_K\}$, where $1 < k < K$. A partition Y of X defines the set of classes which partition X into M groups: $Y = \{y_1, ..., y_m, ..., y_M\}$, where $1 < m < M$. The number of elements counted in a set A is denoted as $|A|$, respectively the number of clusters in C are $|C|$ and the amount of objects in cluster c_k is $|c_k|$. Also, the centroid of a cluster c_k is represented as $\bar{c_k}$.

Moreover, the terms S_C, C_C, V_C, O_C, and D_C represent the separation, cohesion, variance, overlap, and density terms of partition C, respectively. Separation provides a measure of how distinct clusters of a partition are, cohesion and variance estimate how close elements of a cluster are with each other, overlap expressed the degree in which elements of a partition belong to more than one clusters, and density indicates the density of a partition.

2.4.2 External Validation Indices

Homogeneity

The *Homogeneity measure* (Rosenberg and Hirschberg 2007) is estimated by the normalized conditional entropy of a class distribution given a partition of clusters C and classes Y, *i.e.* $H(Y|C)$. The Homogeneity measure indicates if all the data points, which are members of a

given class, are elements of the same cluster. The Homogeneity measure is defined as:

$$
\begin{aligned}
h(Y,C) &= \begin{cases} 1 & \text{if } H(Y,C)=0 \\ 1-\frac{H(Y|C)}{H(Y)} & \text{otherwise} \end{cases} \\
&\text{where,} \\
H(Y|C) &= -\sum_{c_k\in C}\sum_{y_k\in Y}\Big(\frac{|c_k\cup y_k|}{n}\cdot log\frac{|c_k\cup y_k|}{\sum_{y_k\in Y}|c_k\cup y_k|}\Big) \\
H(Y) &= -\sum_{y_k\in Y}\Big(\frac{\sum_{c_k\in C}|c_k\cup y_k|}{|C|}\cdot log\frac{\sum_{c_k\in C}|c_k\cup y_k|}{|C|}\Big)
\end{aligned}
\tag{2.1}
$$

The Homogeneity measure is bounded between 0 and 1, with higher values suggesting a better homogeneous score. Thus, if all data points of distinct classes were assigned to a different cluster then the Homogeneity score would be 1, otherwise, if all data were contained into a single cluster then the measured value would be 0.

Completeness

The *Completeness measure* (Rosenberg and Hirschberg 2007), opposite to Homogeneity, computes the conditional entropy of the partition of clusters C distribution given the partition of classes Y, *i.e.* $H(C|Y)$. The Completeness measure is expressed as:

$$
\begin{aligned}
c(Y,C) &= \begin{cases} 1 & \text{if } H(C,Y)=0 \\ 1-\frac{H(C|Y)}{H(C)} & \text{otherwise} \end{cases} \\
&\text{where,} \\
H(C|Y) &= -\sum_{y_k\in Y}\sum_{c_k\in C}\Big(\frac{|c_k\cup y_k|}{n}\cdot log\frac{|c_k\cup y_k|}{\sum_{c_k\in C}|c_k\cup y_k|}\Big) \\
H(C) &= -\sum_{c_k\in C}\Big(\frac{\sum_{y_k\in Y}|c_k\cup y_k|}{|C|}\cdot log\frac{\sum_{y_k\in Y}|c_k\cup y_k|}{|C|}\Big)
\end{aligned}
\tag{2.2}
$$

Similar to the Homogeneity, the Completeness measure is bounded between 0 and 1, with higher values indicating a better completeness score. Hence, a partition with all data points of a single class assigned to a single cluster is a perfect complete case with Completeness score to 1, whereas if these data points were to be assigned to distinct clusters, then the completeness value would be 0.

V-measure

V-measure score (Rosenberg and Hirschberg 2007) is an external entropy-based cluster validation measure. It is computed as the weighted harmonic mean of the homogeneity and completeness scores. The V-measure score is defined as:

$$V(C,Y) = \frac{(1+\beta)\cdot h(Y,C)\cdot c(Y,C)}{\big(\beta\cdot h(Y,C)\big) + c(Y,C)} \tag{2.3}$$

where β controls the affection Homogeneity and Completeness have over the V-measure score. If $\beta > 1$ the Completeness value influences more strongly the calculation, if $\beta < 1$ then the Homogeneity value has a greater impact, whereas if $\beta = 1$ then both the Completeness and the Homogeneity have equal contribution on the calculation of the V-measure score.

Purity

The *Purity measure* (Ying Zhao and Karypis 2001) provides a metric of homogeneity of a single class across all clusters, and is defined as:

$$P(C,Y) = \sum_{c_k \in C} \Big(\frac{1}{|C|} \cdot max_{y_k \in Y}(|c_k \cup y_k|) \Big) \tag{2.4}$$

Entropy

The *Entropy measure* (Ying Zhao and Karypis 2001), describes how the data labels are distributed across the clusters. This evaluation measure is defined by the summation of the individual cluster entropies weighted with the cluster size, and is defined as:

$$E(C,Y) = \sum_{c_k \in C} \left(\frac{|c_k|}{|C|} \cdot \Big(-\frac{1}{log|Y|} \cdot \sum_{y_k \in Y} \big(\frac{|c_k \cup y_k|}{|c_k|} \cdot log \frac{|c_k \cup y_k|}{|c_k|} \big) \Big) \right) \tag{2.5}$$

Since the *purity* metric is a special version of the *homogeneity* score, and the *entropy* captures the same notion as the *V-measure* score, the *homogeneity*, *completeness*, and *V-measure* metrics are used to evaluate clustering outputs in the following chapters.

Chapter 3

Qualitative Depth-informed Spatial Relations

3.1 Introduction

Qualitative relations and formulae, which are composed of relation symbols, are used to represent semantically meaningful properties of a perceived *event*, *e.g.* interaction, deformation, whilst abstracting away from any non-relevant information, *e.g.* pixel values, numeric measurements. Spatial relations can provide information about the following properties of entities, *i.e.* objects: mereotopology, direction, distance, size, shape, and moving objects (J. Chen et al. 2015; Dylla et al. 2017; Landsiedel et al. 2017).

Inspired by the RCC8 spatial relations and the Process-Grammar (Leyton 1988), the Depth-informed Spatial Relations (DiSR) is a novel set of qualitative spatial relations that exploit the concave regions of the entities in a scene. This new set of relations defines more complex and concave-informed relations, such as "contain" and "support". These relations add expressivity to the purely mereotopological relations of RCC8 and as Chapter 5 will show they are useful in inferring affordances of objects.

This chapter presents a literature review on QSRs in Section 3.2 before introducing the DiSR relations in Section 3.3, and presenting their evaluation in Section 3.4, by assessing their *expressiveness* and *coherence*. Finally, this chapter concludes with Section 3.6.

3.2 Literature Review

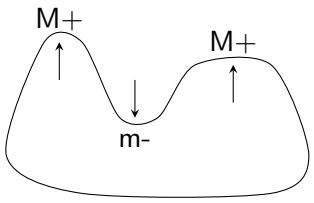
The Region Connection Calculus (RCC) (Anthony G Cohn et al. 1997; Gerevini and Renz 2002) and the n-intersection model (Egenhofer 2005; Egenhofer and Franzosa 1991; Egenhofer and Franzosa 1995; Egenhofer and Herring 1990; Egenhofer and Sharma 1993; Egenhofer and Vasardani 2007; Egenhofer, Clementini, et al. 1994; Egenhofer, Mark, et al. 1994) are the most well-known and well-studied approaches for representing topological information. However, as both sets of relations solely rely on the spatial location of the detected entities, *i.e.* objects, occlusion cannot be differentiated from an actual interaction of containment. Galton's *Lines of Sight* (Galton 1994), the Region Occlusion Calculus (D. Randell et al. 2001), and The Occlusion Calculus (OCC) (Köhler 2002) aim to solve the occlusion problem in 2D space. Moreover, the *Depth Profile Calculus* (Santos 2007) defines a set of qualitative relations aiming to detect occlusions of objects based on a sequence of depth profiles, *i.e.* depth information acquired by visual sensors of a scene. However, relations such as "contain" are still not representable. Thus, shape characteristics are crucial for detecting more relevant relations to object interactions.

Qualitative shape information can be obtained by using region-based methods (Anthony G Cohn 1995) or boundary-based methods (Leyton 1988; Galton and Meathrel 1999; Gottfried 2003b; Gottfried 2003a; Gottfried 2004). Both kind of methods depend on n-D detected entities. SpOKE (Bennett et al. 2013) comprise of various spatial relations that take into consideration the mereotopological, as well as the shape information of entities. In this qualitative relational set, the containment relation is expressed by the interactions with a detected cavity of an object. However, to represent a possible containment relation in 3D space, the cavity information needs to be detected from visual data.

3.3 Depth-informed Relations

Though simple and discrete spatial interactions, *e.g.* "touching" and "not touching" in 2D space, are captured effectively in the 2D plane, the determination of more complex spatial relationships, *e.g.* "supporting", "containing", is challenging, especially when considering a cluttered scene. The Depth-informed Spatial Relations (DiSR) set of qualitative spatial relationships addresses this limitation, enabling the distinction between occlusion and interactions, and reasons about the spatial relative position of objects in 3D space, by taking into account the object's convexity type.

To detect the convexity type of objects in the scene, one must obtain knowledge about the *indentation* (m-) and *protrusion* (M+) areas of an object, as defined in the *Process-Grammar* (Leyton 1988). 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+, as illustrated in the schematic below.



Such information is critical to ascertain a relation when two objects interact in the 2D plane, *e.g.* a "containment" (Cont, Conti) DiSR relation occurs between one or more "concave" type objects when the "containee" confirms that is between the m- and M+ areas of the "container".

The detection of the *indentation* and *protrusion* areas of an object depends on the means of input data. In point cloud data *i.e.* 3D, one can effectively detect and segment these areas of the objects in a scene by considering the third dimension for every data point of an object. However, on image data the detection of the convexity type of the objects relies on depth cues. Thus 2.5D information is taken into account, by considering the RGB and Depth information captured from monocular-infrared sensor pairs. The proposed method is dependent on the viewing angle of the camera, as well as occlusions of possible indentation areas due to orientation of the objects captured.

Convexity type detection with 2.5D data

Three primary convexity type objects are identified: *concave*, *convex*, and *surface*. For the determination of the objects' convexity type the increasing-ordered-by-value depth information across the pixels defining each object is considered ($\text{dist}_{\text{depth}}$) (Figure 3.1(right)).

Algorithm 1 is employed for identifying the objects convexity type, and it is based on a convexity depth threshold ($thresh_{convex}$), which defines the upper boundary of depth range information of a "concave" or "surface" type object.

In this work, an absolute value is considered for the convex depth threshold. This constrains the method to detect only known concavities based on the data that were considered for tuning

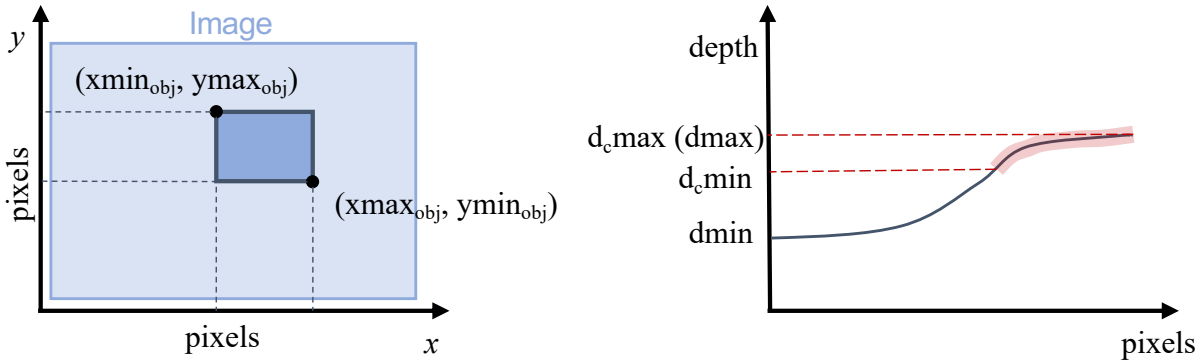


Figure 3.1: (left) Detected object bounding box coordinates. (right) Object depth distribution and the depth information signifying potential concave curve detection between the values $d_c min$ and $d_c max$ (red).

Algorithm 1 Compute the convexity type of an object.

Given: $thresh_{convex}$

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

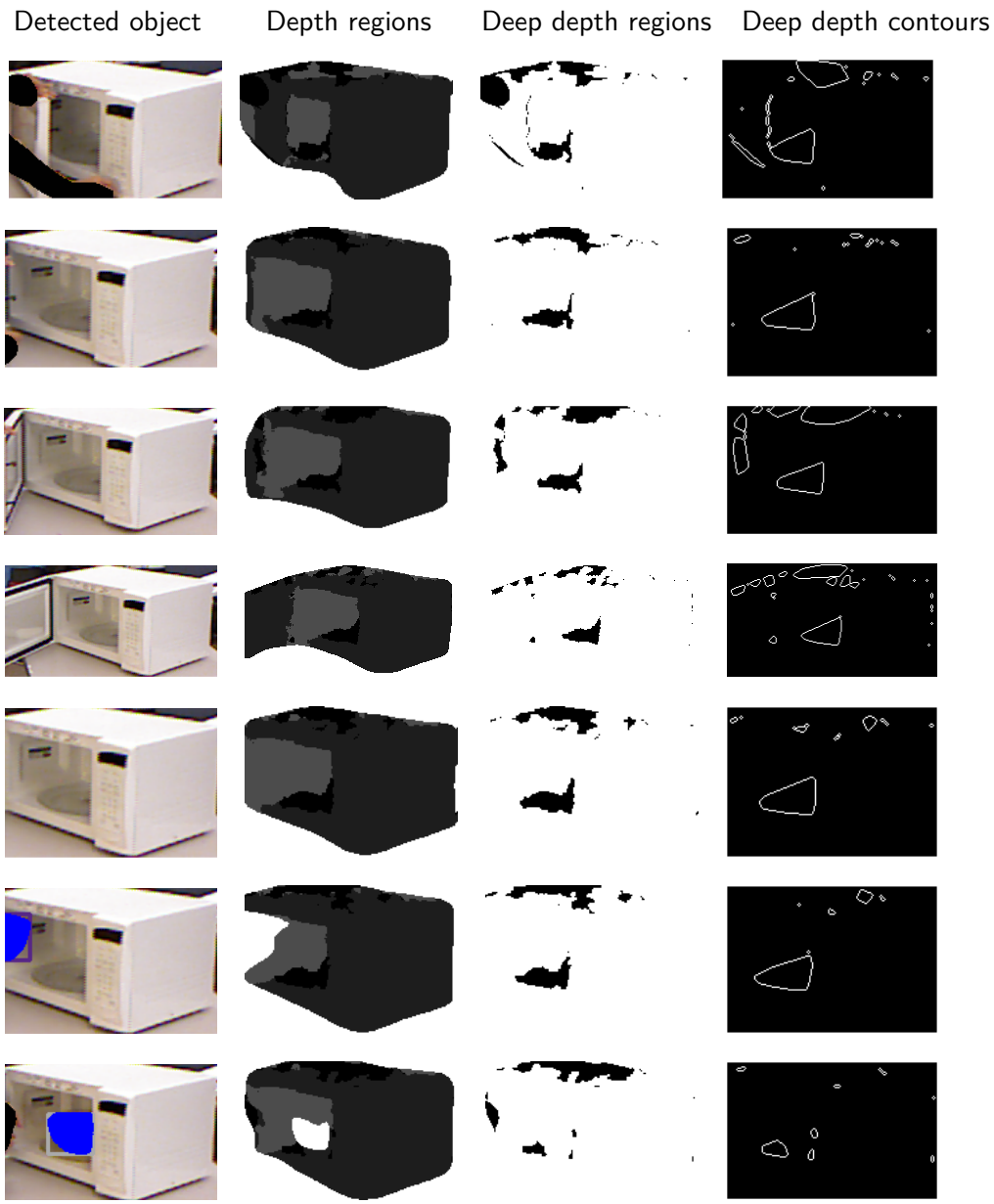


Figure 3.2: Processing steps for extracting depth contours from 2.5D data.

that threshold. A more generalizable approach would be to employ the relative shape/size of the object as well as its distance from the camera, for considering a dynamic convex depth threshold.

Objects with depth range (considering the minimum and maximum values of the $\text{dist}_{\text{depth}}$) 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 $thresh_{convex}$. Figure 3.2 illustrates the processing steps for extracting depth contours. The bounding boxes or

masks of the detected objects in the RGB frames are mapped onto the depth data. The depth information corresponding to every object is then partitioned into sections depending on their depth values. Finally, from that depth data, deep regions are considered from which depth contours are created indicating potential *indentation* areas, thus classifying objects as "concave". Any information indicating that a detected object is occluding another one, the information of the first object is discarded from the latter; in Figure 3.2 the last two rows illustrate the presence of a bowl object (blue) whose depth information is not considered as being part of depth data of the microwave, as shown in the "Depth regions" column.

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. The frame of the image is also a contour, and it is the root of the tree. Contours which are not directly included in the root contour are not considered its children, thus are not considered as potential indentention areas. Potential contour noise can be eliminated by considering the relative difference of the size of the contours with respect to the detected object. The remaining contours are then examined in terms of hierarchy. A *child contour* indicates an *indentation area.* Thus, the detection of a child contour in the depth domain indicates the presence of a concave curve, therefore a "concave" type object, and "surface" type otherwise [1]. The process of the detection of a concave object is summarized in Figure 3.3, in which from an RGB video frame, objects are detected using an off-the-shelf object detector. The bounding boxes or masks proposing an object detection are mapped onto the depth domain to extract the depth data relevant to that object detection. The depth data is then partitioned and deep depth regions are exploited to produce depth contours of possible concave curve areas. Finally, depth contours are a hierarchical set depending on each contour's inclusion, *e.g.* the orange contour is included in the red contour, hence it is a child of the red contour. If a child contour exists then the region it encapsulates is detected as a concave curve area. Moreover, if more than one child contour is detected, the larger one is considered as a concave curve area.

Potential concave curve detection with 2.5D data

Defining only the type of an object is not sufficient for reasoning about the spatial interactions between objects. Detecting complex spatial object relations effectively, *e.g.* "contain", is

[1] This approach considers this restricted representation for detecting concave objects as the depth resolution varies depending on the objects' distance from the visual sensor.

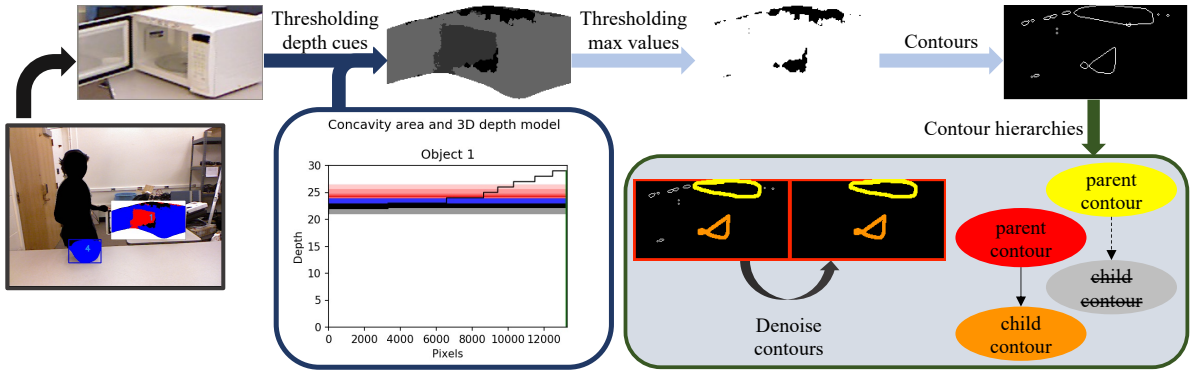


Figure 3.3: Process steps for detecting "concave" type objects with 2.5D data.

challenging without considering potential concave curves of objects. Thus, Algorithm 2 is used to infer the boundaries of the m- and M+ areas of an object, as a direct generalization of the *Process-Grammar* to 2.5D, indicating a concave curve in 3D space, with respect to the object's depth information and convexity type.

For a "concave" type object the ordered-by-value depth information ($\text{dist}_{\text{depth}}$) is partitioned into *sd* sections for distinguishing the *indentation* from the *protrusion* area. The sd_{max} sections with the highest depth values are estimated to capture its concave curve [2]. These depth boundaries of such objects are set to enclose the potential concave curve's depth information for detecting the relation "contain" (Cont, Conti). Depth boundaries of "convex" and "surface" type objects are not processed due to concave curve absence. d_cmax and d_cmin set the boundaries of the potential concave curve area of a "concave" type object and are further employed for detecting a "contain" relation between a pair of objects.

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

Given: sd, sd_{max}

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

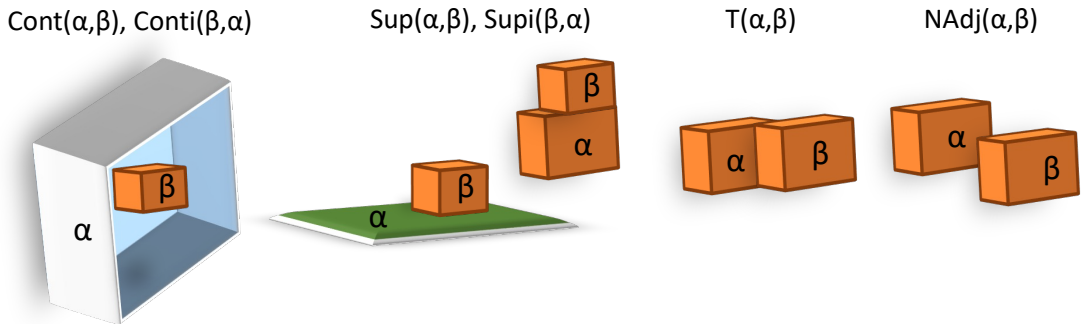


Figure 3.4: DiSR representations in 3D space.

3.3.1 Formulation of *DiSR*

The DiSR set comprise of the relations[3]: "supports" (Sup, Supi), "contains" (Cont, Conti), "touching" (T), and "not adjacent" (NAdj). Figure 3.4 illustrates the DiSR representations in 3D space. For a spatial interaction to hold between a pair of objects in the scene, a depth distribution overlap must be evident. Depending on the kind of depth information overlap and the convexity-type of the interacting objects, a different qualitative spatial relation from the DiSR set holds. The detection of which relation holds every time relies on the following diagram scheme.

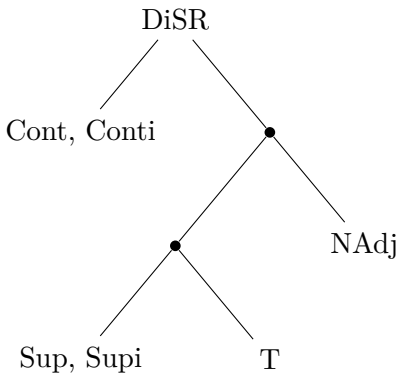


Table 3.1 presents the definition of every DiSR relation inspired by the RCC set, whilst exploiting the RCC relations: "overlap" (O) and "part" (P, Pi) computed in the 2D image plane. Moreover, it employs the relations DPO, DPP, DPPi and On, which are further defined in this Section.

"Depth overlap" (DPO) and "depth proper part" (DPP, DPPi) are primitive relations which

[2]The parameterization of sd, sd_{max}, and $thresh_{convex}$ is conducted in an empirical study and is further explained in Section 5.5.1.

[3]The inverse of a relation is denoted with an i at the end, wherever an inverse exists.

hold between objects' depth information, defined as:

$$\mathsf{DPO}(\alpha,\beta) \equiv ((dmax_\alpha \geq dmin_\beta) \wedge (dmax_\alpha < dmax_\beta) \wedge$$
$$(dmin_\alpha < dmin_\beta)) \vee ((dmax_\beta \geq dmin_\alpha) \wedge$$
$$(dmax_\beta < dmax_\alpha) \wedge (dmin_\beta < dmin_\alpha))$$
$$\mathsf{DPP}(\alpha,\beta) \equiv (dmax_\alpha > d_cmin_\beta) \wedge (dmax_\alpha \leq d_cmax_\beta) \wedge$$
$$(dmin_\alpha \geq d_cmin_\beta) \wedge (dmin_\alpha < d_cmax_\beta)$$

where $dmax$ and $dmin$ are the maximum and minimum depth values, respectively, by considering the depth cues of the detected object. Also, d_cmax and d_cmin are the maximum and minimum depth values of the concave region (if detected), respectively, derived from Algorithm 2.

Moreover, the spatial relation On is defined in the 2D space as,

$$\mathsf{On}(\alpha,\beta) \equiv \mathsf{O}(\alpha,\beta) \wedge ((ymax_\alpha \geq ymax_\beta) \wedge$$
$$(ymin_\alpha \geq ymin_\beta) \wedge (xmax_\alpha \leq xmax_\beta) \wedge (xmin_\alpha \geq xmin_\beta))$$

where $(xmin, ymax)$ and $(xmax, ymin)$ are the top-left and bottom-right corners of the detected object's bounding box, respectively (Figure 3.1(left)).

3.4 Evaluation

The evaluation of the DiSR set comprises of two evaluation tasks[4]:

1. evaluating how well the DiSR set can describe and distinguish object interactions in the real-world (*expressiveness*), and;
2. evaluate how coherent the DiSR set is (*coherence*), *i.e.* their cognitive validity.

The experiments were conducted in an online 3D virtual environment[5] with human participants, who had to interact with the objects in the scene. In total, 27 humans participated in both experimental tasks. The requirements for recruiting these participants were: to be at least 18

[4]The study has been approved by the University of Leeds Ethics committee (ref LTCOMP-008, date of approval 09 November 2022).

[5]`https://www.vectary.com`

DiSR relations		
Relation	**Definition**	**Description**
$\mathsf{Cont}(\alpha, \beta)$	$\mathsf{P}(\beta, \alpha) \wedge \text{Concave}(\alpha) \wedge \mathsf{DPP}(\beta, \alpha)$	α contains β
$\mathsf{Sup}(\alpha, \beta)$	$\mathsf{DPO}(\alpha, \beta) \wedge (\text{Surface}(\alpha) \vee \mathsf{On}(\beta, \alpha))$	α supports β
$\mathsf{T}(\alpha, \beta)$	$\mathsf{O}(\alpha, \beta) \wedge \mathsf{DPO}(\alpha, \beta) \wedge \neg\mathsf{Cont}(\alpha, \beta) \wedge \neg\mathsf{Cont}(\beta, \alpha) \wedge \neg\mathsf{Sup}(\alpha, \beta) \wedge \neg\mathsf{Sup}(\beta, \alpha)$	α is touching β
$\mathsf{NAdj}(\alpha, \beta)$	$\neg\mathsf{Sup}(\alpha, \beta) \wedge \neg\mathsf{Sup}(\beta, \alpha) \wedge \neg\mathsf{Cont}(\alpha, \beta) \wedge \neg\mathsf{Cont}(\beta, \alpha) \wedge \neg\mathsf{T}(\alpha, \beta) \wedge \neg\mathsf{T}(\beta, \alpha)$	α is not adjacent to β
C, DC, P, and O of RCC relations		
$\mathsf{C}(\alpha, \beta)$ ‡	$mask(\alpha) \cap mask(\beta) \neq \varnothing$ †	α and β are connected
$\mathsf{DC}(\alpha, \beta)$	$\neg\mathsf{C}(\alpha, \beta)$	α is disconnected from β
$\mathsf{P}(\alpha, \beta)$	$\forall z(\mathsf{C}(z, \alpha) \rightarrow \mathsf{C}(z, \beta))$	α is a part of β
$\mathsf{O}(\alpha, \beta)$	$\exists z(\mathsf{P}(z, \alpha) \wedge \mathsf{P}(z, \beta))$	α overlaps β

† $mask(x)$ defines object x's region in the image plane, that represents a collection of pixels.
‡ If either α or β comprise a single pixel, then C holds if a pixel from α and a pixel from β are adjacent. Whereas, if α and β comprise of more than 2 pixels, then C holds when there is at least one pixel shared between them.

Table 3.1: DiSR formulations based on the RCC relations.

years old and to speak English. There were no restrictions on whether the participants were native English speakers or not[6].

Participants were recruited through email, using several emailing lists of the University of Leeds, as well as through in-person contact. A compensation was offered (entering a draw to win a £50 Amazon voucher) to everyone who took part in these experiments and did not wish to withdraw. At the start of the experiments, participants were given an information sheet and an instruction sheet (Appendix Section B.1) which they were asked to read before commencing. They could ask any question before starting the experiments to better understand the tasks they had to complete. Moreover, a demonstration of how to use the online 3D virtual environment was provided.

3.4.1 Evaluating DiSR Expressiveness

For the evaluation of the *expressiveness* of the DiSR, a set of 16 predefined synthetic scenes presenting a pair of interacting objects, was given to the participants to collect human data of DiSR detections. Figure 3.5 presents these synthetic scenes. The human participants could rotate around the configured objects to have a better understanding of their interaction, and

[6]All participants had a student/staff property in a university in the United Kingdom. Hence, they were considered to have adequate knowledge of the English language considering the corresponding admittance criterion of the university.

were asked to select one of the DiSR relations (Cont, Sup, T, NAdj) for each pair of objects.

A K-means clustering of that data is presented in Figure 3.6, where the y-axis represents the class labels of the interactions ("support", "notadjacent", "contain", "touching") according to the proposed algorithm, and the x-axis visualizes the clusters' id ("0", "1", "2", "3"). The figure illustrates for every cluster produced the distribution of data in reference to the relation labels. The sharper the distributional plot the more homogeneous and complete the cluster of the specific label. It is evident that the relations "notadjacent" as well as "contain" are well distinguishable from the rest as a single and sharp peak is presented for these labels. However "touching" and "support" are not clearly separated in terms of English meaning, since more than one peaks are formed, causing the "support" relation to be present in a significant percent in cluster "2", which is dominated by the "touching" relation, and the "touching" relation to be also present in cluster "1", where it is mostly comprised of the relation "support". The clustering produces groups of data with homogeneity score reaching 74%, completeness score of 75%, and v-measure 74%.

Figure 3.7 illustrates for every relational category in the DiSR set the percentage of the detected relations from the human participants.

3.4.2 Evaluating DiSR Coherence

For the evaluation of the *coherence* of the DiSR set, 16 DiSR relations between pairs of objects were given to human participants and were asked to create in the 3D virtual environment the interaction described by the DiSR relation in the scene, illustrating how the objects would be configured in the real-world for every relation to hold.

The DiSR relations given to be configured in the 3D scene are presented in Table 3.2.

16 scenes were initialized to contain the relevant objects. By moving, rotating and resizing them in the 3D environment the participants were able to configure them accordingly. To evaluate how well the DiSR relations are comprehended, a qualitative study is performed by considering the object configurations from the human participants for the defined scenes from Table 3.2. Figure 3.8, Figure 3.9, Figure 3.10, and Figure 3.11 illustrate the representative instances[7] of these qualitative results for all the defined scenes.

[7]This is not a complete visualization of the data captured. These images showcase the diversity of object configurations from the data acquired form the human participants.

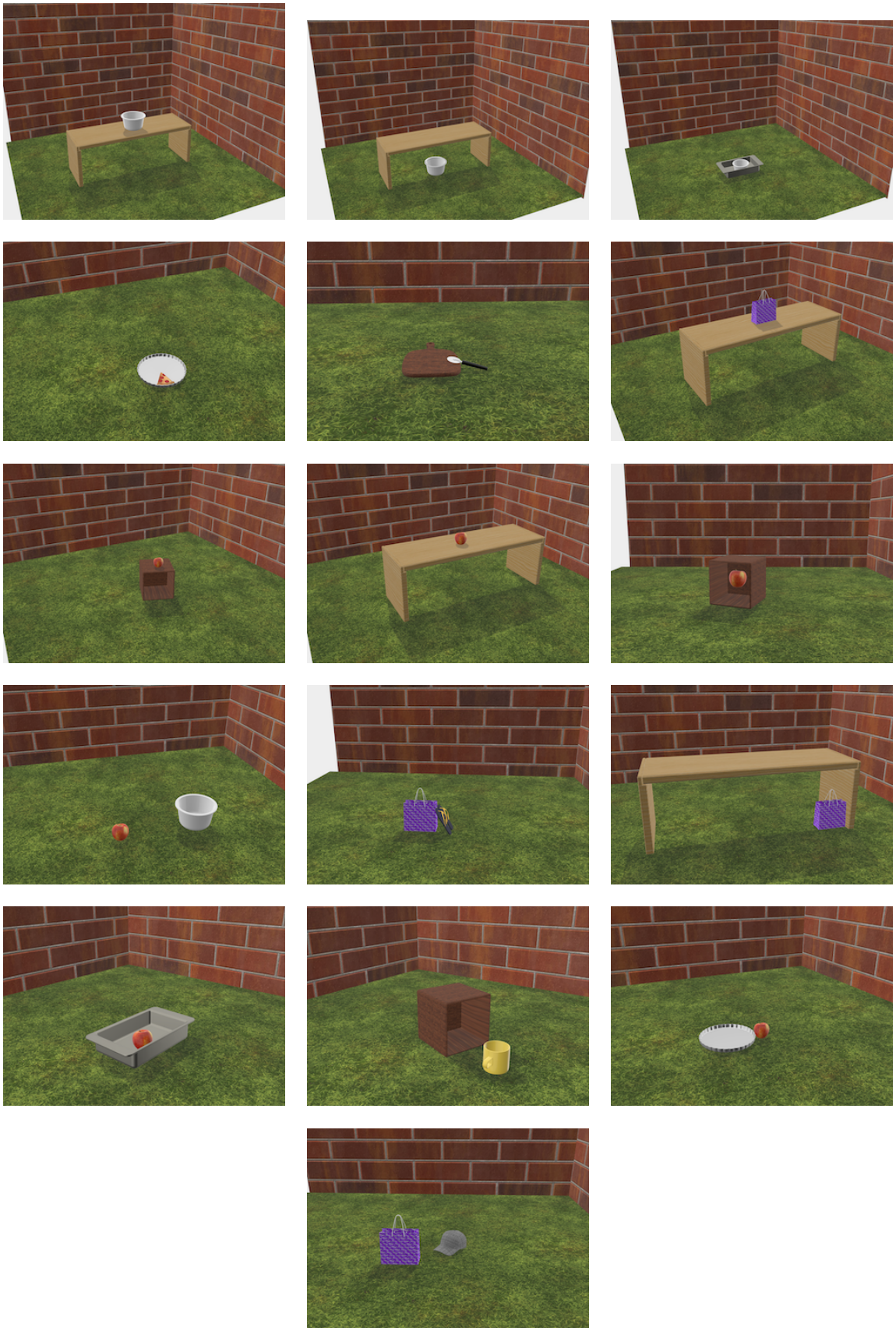


Figure 3.5: 3D virtual scenes to evaluate expressiveness of the DiSR relations.

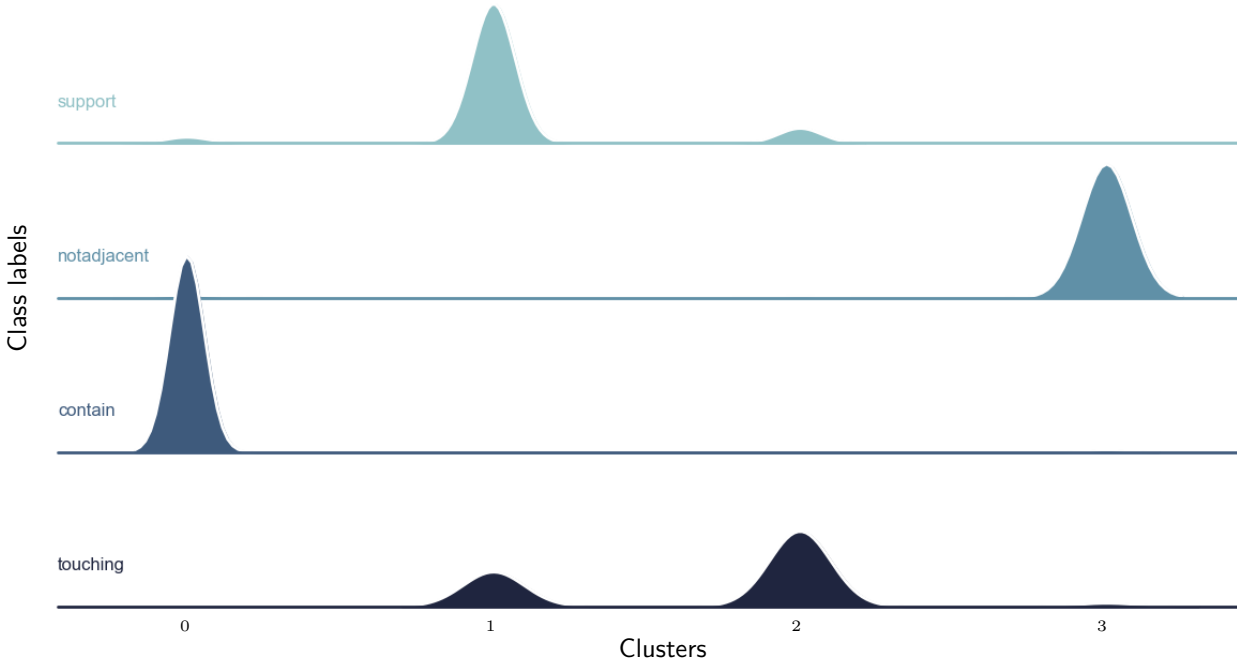


Figure 3.6: The DiSR detections from the collected human data are clustered using K-means, forming 4 groups of relations, *i.e.* clusters 0, 1, 2, and 3. The class labels are shown on the y-axis of the plot, presenting the distribution of every class label into each of the clusters. The sharper the peak formed the more complete and homogeneous the cluster for a specific class label is.

Scene ID	DiSR
1	Sup(bench, backpack)
2	T(backpack, bench)
3	NAdj(table, armchair)
4	T(bin, table)
5	Sup(armchair, book)
6	Cont(baking dish, bread)
7	Cont(plant pot, ball)
8	NAdj(umbrella, chair)
9	T(tree, leaf)
10	Sup(shoe box, shoes)
11	NAdj(street lamp, bench)
12	Sup(pot, pot holder)
13	Cont(sink, mug)
14	T(box, ball)
15	NAdj(skateboard, shoes)
16	Cont(tent, car)

Table 3.2: DiSR relations given to evaluate coherence for every scene. Each scene comprises the relevant objects and the participants are asked to configure the objects appropriately so the corresponding DiSR relation holds.

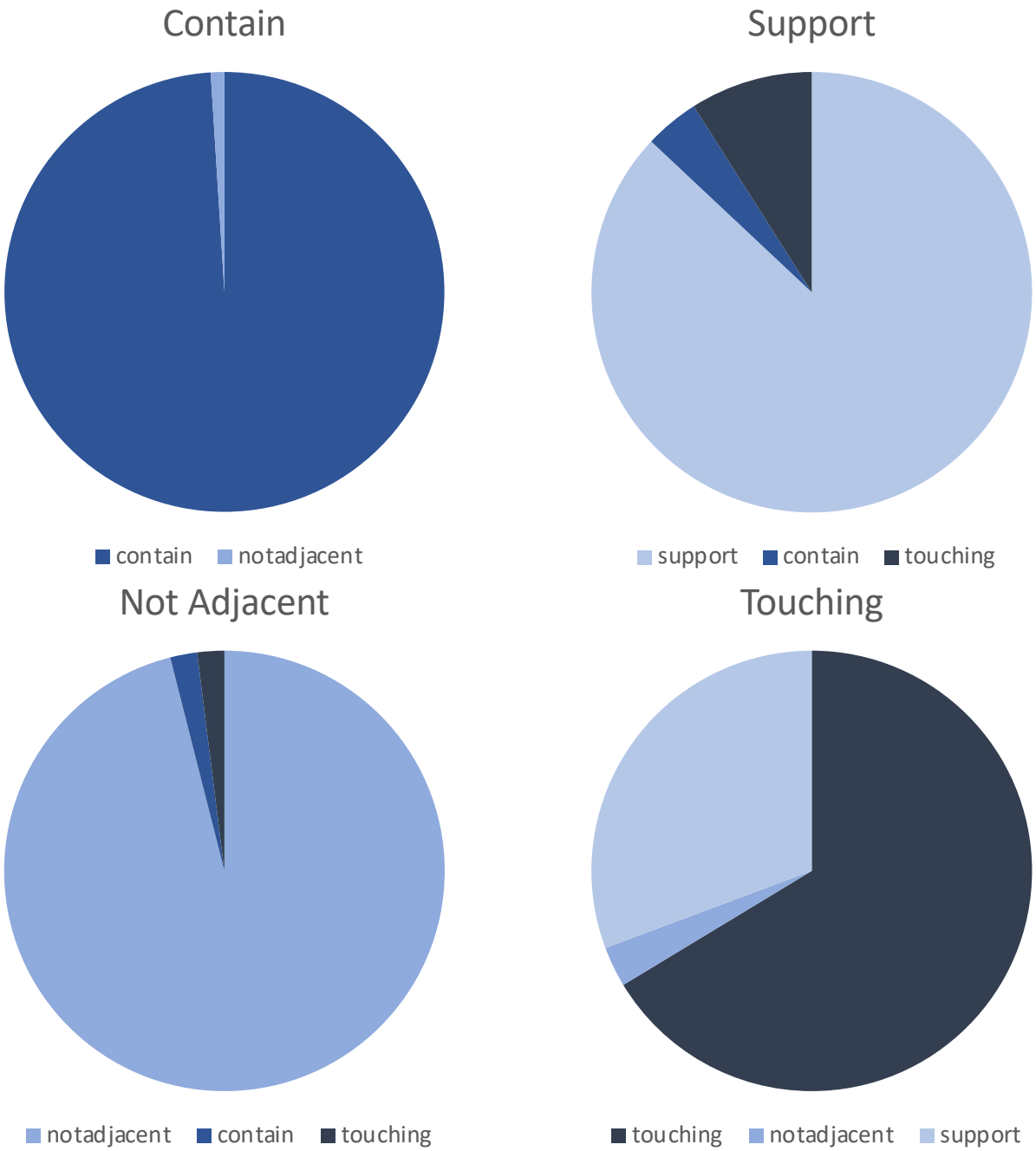


Figure 3.7: Percentage of DiSR detections for every DiSR relation. The label of every disc plot notes the ground truth relation, and the relations within each disc plot show the appointed relations from the collected human data.

From the collected data, it is evident that the participants didn't always exploit the surface of an object to configure objects for a "support" relation to hold, as it was indicated to them. *E.g.* in scenes 2, 5, 10 and 12, some participants visualized the "support" relation as "partial support", *i.e.* an object leaning on another object. This is due to the fact that the definitions provided for the DiSR relations are an approximation of the English meaning they refer to, thus an inherit human bias of the English meaning of the words cause these false negative and false positive DiSR configurations.

A quantitative analysis visualized in Figure 3.12 presents the percentages of every DiSR relation detected from the participants data in each scene. The DiSR definitions are used for each participant's configuration to determine the relationship holding between the corresponding pair of objects. From these column plots it is apparent that whenever a "support" relation occurs, 10-20% of the participants provide object configurations where "touching" relation holds instead, *e.g.* in scenes 1, 5, 10, and 12 where a "support" relation originally holds, around 10-20% of the participants configured the objects for a "touching" relation.

3.5 Limitations

This Chapter proposes the DiSR set, which comprise of definitions of relations that are an approximation of the English meaning they are referring to. *E.g.* the word "support" in English is employed to describe the event of an object holding another object at a specific height, *e.g.* the table supports the bowl, the nail supports the frame, or an object holding another object from falling over, *e.g.* the book holder supports the books. Moreover, "contain" in English can describe different kind of containment, *i.e.* an object being within the cavity of another object, *e.g.* the bowl is contained in the microwave, or an entity being part of another entity, *e.g.* the drink contains sugar.

There can be situations where the shape of a "concave" type object, allows other objects to interact with a non-concave area, with similar depth information as in the cavity region. The proposed approach is limited to objects which do not have non-concave regions with similar depth as in the cavity area. An enhancement of the DiSR definition of Cont can address this limitation by considering the location of the concave region.

Moreover, the On relation is defined in 2D space, not allowing the generalization across a wide

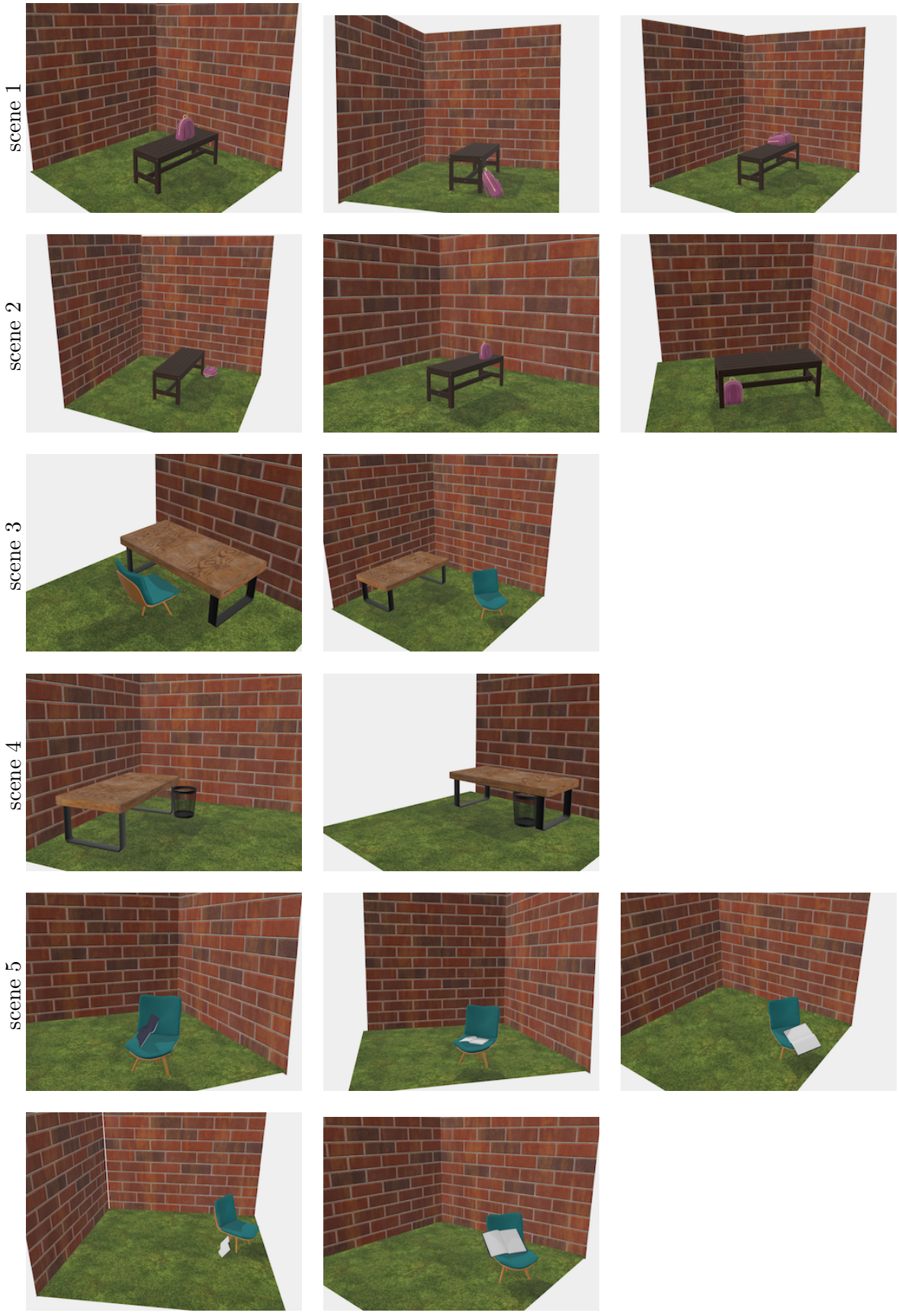


Figure 3.8: Illustration of DiSR interactions from human agents (scenes 1 to 5).

Figure 3.9: Illustration of DiSR interactions from human participants (scenes 6 to 10).

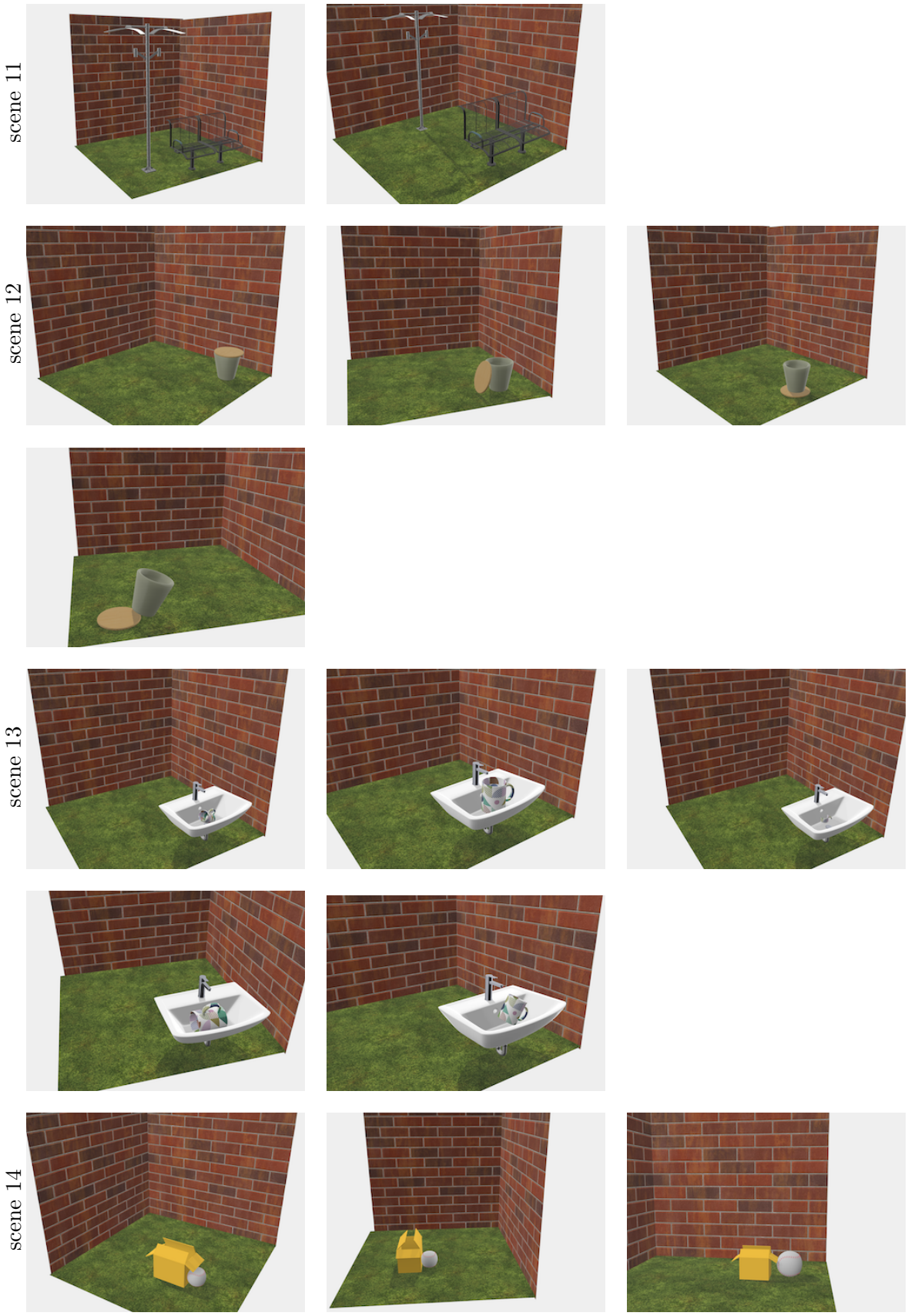


Figure 3.10: Illustration of DiSR interactions from human participants (scenes 11 to 14).

Figure 3.11: Illustration of DiSR interactions from human participants (scenes 15 to 16).

variety of actual "ontop" situation to be detected, *e.g.* wearing a hat *on* the head where the sides of the hat exceed the head diameter.

Furthermore, the detection of the DiSR relations relies on the camera view point. For a Cont relation to hold, it is necessary to detect the concave curve on the object of interest, thus if a containment holds but it is not visible to the visual sensor, the Cont relation cannot be detected.

In this work, an absolute value is considered for the convex depth threshold. This constrains the method to detect only known concavities based on the data that were considered for tuning that threshold. A more generalizable approach would be to employ the relative shape/size of the object as well as its distance from the camera, for considering a dynamic convex depth threshold.

3.6 Conclusions

The qualitative DiSR set comprises of a set of qualitative spatial relations. These relations take into account the shape of the detected entities, to detect more relevant relations between interacting objects, *e.g.* surface objects are able to support other objects. The definitions proposed are an approximation of the English meaning they are referring to: it is possible to satisfy the definitions by configurations which do not 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

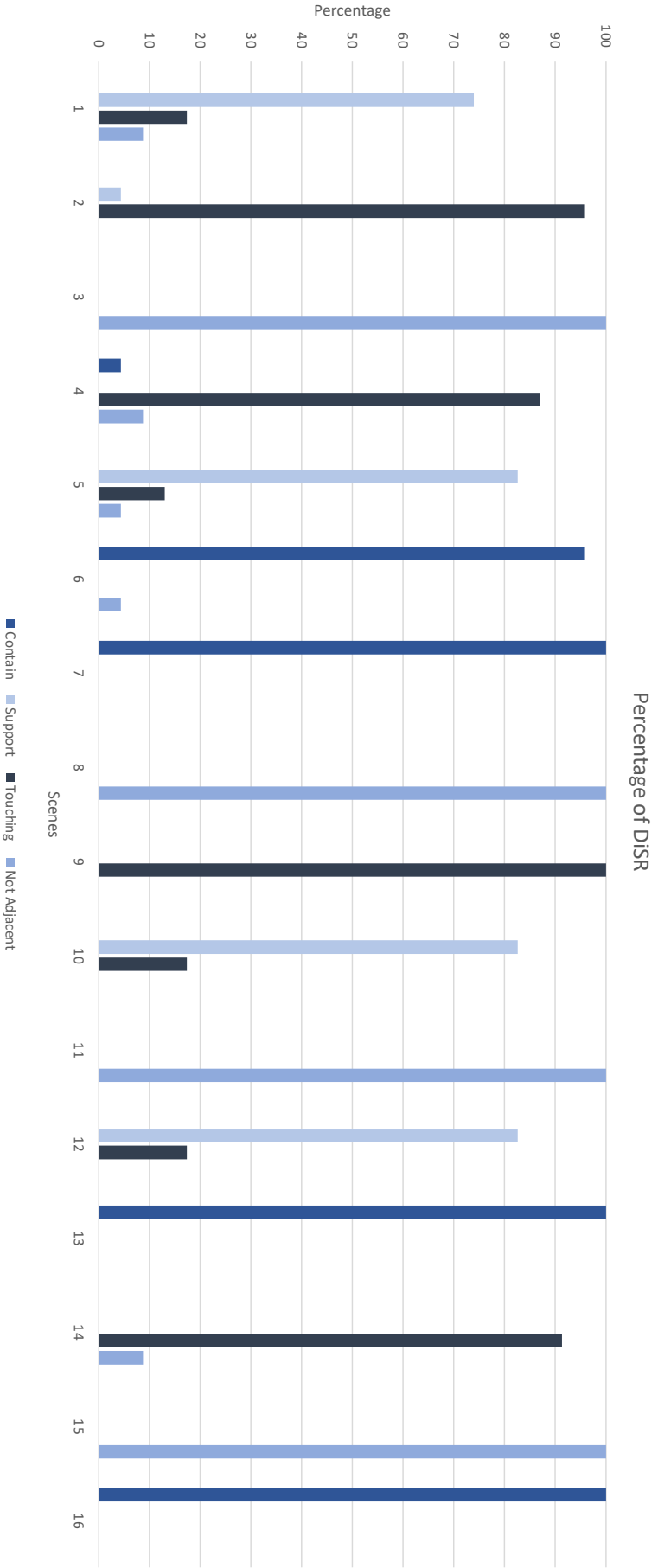


Figure 3.12: Percentage of each DiSR relation on every scene for studying the coherence.

middle object in between. Nevertheless, these definitions are easy to compute and work well in the everyday-life scenes considered so far.

Chapter 4

Leeds Object Affordance Dataset

4.1 Introduction

Performing inference on object interaction data, acquires the use of input data that provides knowledge about the objects' spatial as well as temporal interactions. *E.g.* the interaction of stirring a spoon in a cup needs the temporal information to be distinguished from only containing the spoon in a cup. Hence, a sequence of image frames, *i.e.* video data, capturing human-object interactions should be exploited for these tasks.

The task of object affordance prediction/categorization relies on object interaction data. The general notion of an affordance of an object is considered the *role* it holds in an interactions. *E.g.* a microwave can play the role of a container when it is captured to be containing another object, but it can also act as a supporter when it supports another object. Hence, object affordances are correlated with the way objects are being used in an activity and their interaction with other objects. Moreover, an object can have multiple affordances, *i.e.* can play different roles, depending on the interactions taking place. Chapter 5 focuses on the categorization of object affordances and properly defines object affordances.

The task of object affordance prediction/categorization relies on object interaction data. Thus, it is crucial for the visual data to capture various object roles. Such data should not be limited to visualize the functionality of an object, *i.e.* the purpose of its creation. However, different utilizations of the objects should be recorded, exploiting its visual characteristics, *e.g.* shape, size.

This chapter presents a review of the literature to date, of available datasets that can be used for tasks such as, object interaction prediction, and affordance prediction and categorization, in Section 4.2. Moreover, the newly publicly available Leeds Object Affordance Dataset (LOAD) [1] is introduced, in Section 4.3, which aims to push the boundaries of affordance distinction and categorization through complex real-world everyday-life activities. Finally, this chapter concludes with Section 4.4.

4.2 Literature Review

From the literature, there are some video datasets that comprise of human activities, and are focusing on human-object interactions rather than human actions and scene semantics. A subset of these datasets are non-egocentric, and are presented in this section.

The MSRDailyActivity3D (J. Wang et al. 2012) and MSRActionPair (Oreifej and Liu 2013) datasets comprise of RGB-D video data of human-object interactions captured with a Kinect sensor. In the MSRDailyActivity3D dataset, 10 subjects perform every activity twice, *e.g.* on the "sitting on sofa" and the "standing" pose, whereas in the MSRActionPair dataset, 10 actors repeat every activity three times. The video data of both datasets represent simple and discrete human-object interactions without any clutter in the scene. However, in the real-world, indoor as well as outdoor scenes appear to be a lot more cluttered with objects and reveal more complex interactions with multiple objects.

The CAD-60 (Sung et al. 2012) and CAD-120 (Koppula et al. 2013) datasets address the limitations of MSRDailyActivity3D and MSRActionPair datasets by capturing RGB-D video data of human-object interactions with multiple objects in cluttered scenes, whilst using a Kinect sensor. They involve 60 and 120 RGB-D videos, respectively, visualizing daily activities in difference scenarios, *e.g.* kitchen, office. Similar to CAD-60 and CAD-120, the Watch-n-Patch (C. Wu et al. 2015) dataset includes RGB-D video data of human-object interactions, however it comprises a larger set of videos.

Different from all the aforementioned datasets, the EPIC-KITCHENS dataset (Damen, Doughty, Farinella, Fidler, et al. 2018; Damen, Doughty, Farinella, Fidler, et al. 2021; Damen, Doughty, Farinella, Furnari, et al. 2022) is a large-scale RGB-D collection of egocentric video data capturing human-object interactions in-the-wild, related with any kitchen scenario, *i.e.* cooking,

[1] `https://doi.org/10.5518/1186`

Dataset	data	#vid.	avg.fr.[a]	a.p.v.[b]	activities
MSRDaily-Activity3D	2.5D	320	N/A	1	drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up, sit down
MSRAction-Pair	2.5D	360	N/A	1	pick up/put down a box, lift/place a box, push/pull a chair/, wear/take off a hat, put on/take off a backpack, stick/remove a poster
CAD-60	2.5D	60	-	1-2	brushing teeth, cooking (stirring), writing on whiteboard, working on computer, talking on phone, wearing contact lenses, relaxing on a chair, opening a pill container, drinking water, cooking (chopping), talking on a chair, and rinsing mouth with water
CAD-120	2.5D	120	525	1-2	arranging objects, cleaning objects, having meal, making cereal, microwaving food, picking objects, stacking objects, taking food, taking medicine, unstacking objects
Watch-n-Patch	2.5D	457	170	1-2	turn on monitor, turn off monitor, walking, play computer, reading, fetch book, put back book, take item, put down item, leave office, fetch from fridge, put back to fridge, prepare food, microwaving, fetch from oven, pouring, drinking, leave kitchen, move kettle, fill kettle, plug in kettle
LOAD	2.5D	58	258	3-4	pull chair, push chair, sit on chair, sit on the floor, sit on the table, push table, pull table, hold bottle, hold umbrella, carry bag, drinking, put item down, kick ball, roll ball, bounce ball, throw ball, cover, put into, take out, lean against the wall, roll bottle, roll chair, roll suitcase, carry chair

[a] average number of frames (per video)
[b] activities per video

Table 4.1: Comparison of datasets capturing human-object interactions in everyday-life activities.

cleaning dishes, stirring, cutting.

Table 4.1 summarizes the key characteristics of all the previously mentioned non-egocentric datasets.

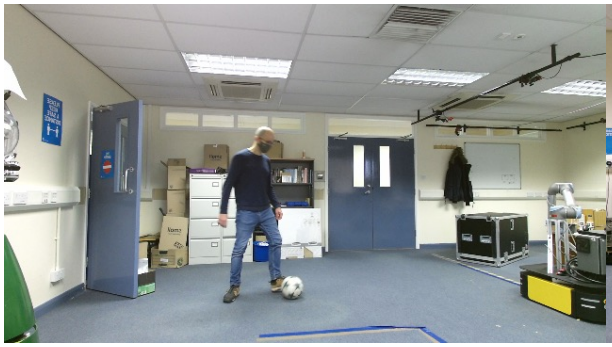
4.3 LOAD Description

The Leeds Object Affordance Dataset or LOAD set effectively extends the domain of activities existing in all the publicly available datasets to date. It comprises of 58 RGB-D videos whilst employing 3 human participants. The video collection captures various human-object interactions in cluttered scenes, which are not present in the existing video datasets, visualizing human-object interactions. The recorded activities were semi-scripted, *i.e.* high-level instructions were given to the human agents prior to the data collection process, *e.g.* "play with the ball", "push/pull the chair and leave the bottle down". Sample frames from the LOAD videos are presented in Figure 4.1.

The activities visualized in the dataset are "play with ball", "drink", "sit", "move objects", "cover", and "lean". Figure 4.2 presents the percentage of these activities within the whole dataset. It is evident that "move objects" and "sit" comprise of the largest portions of the dataset. This is due to the fact that the actions "pull", "push", "roll", "hold", "carry", "put down", "put into", and "take out" are sub-categories of ''move objects", and the "sit" activity contains the sitting action of a human whether using a chair, a box, the floor, or even when they are laying down.

Detectable object affordances in LOAD are: "supportable", "can support", "leanable", "can lean", "containable", "can contain", "coverable", "can cover", "sittable", "holdable", "pullable", "pushable", "rollable", "carriable", "kickable" , "drinkable", and "bouncable". Figure 4.3 illustrates the percentage of the affordance labels occurring in LOAD; area of rectangle reflects abundance in the dataset. In this dataset, apart from the interactive objects, the floor of the scene is also considered as an entity that holds an affordance, thus the affordance labels of "can support" and "supportable" dominate in the ground truth of the affordance labels. Also, it is evident that several affordance labels do not have complementary affordance, *e.g.* "kickable", "pushable", "pullable". This is due to affordance-related interactions caused by the human agent, *e.g.* the human is pushing/pulling a chair, the human is kicking/bouncing a ball.

ball: kickable, supportable

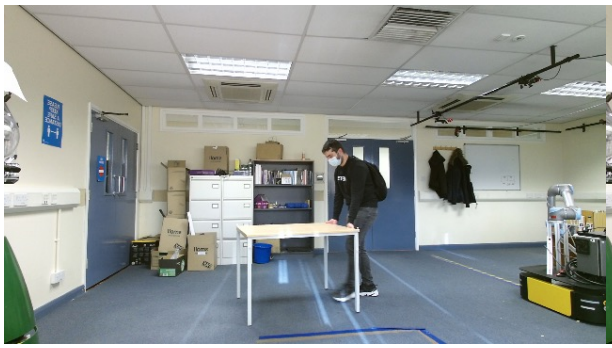


ball: holdable, bouncable

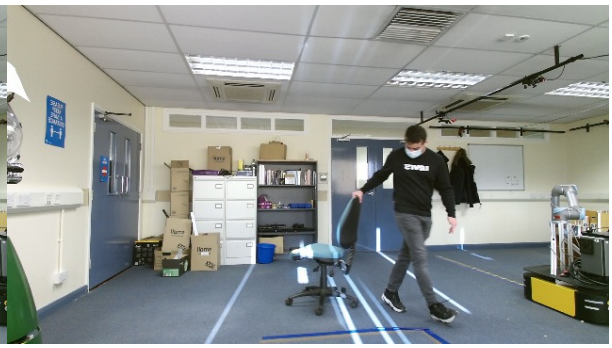
suitcase: holdable, carriable

chair: holdable, carriable

table: supportable, pushable

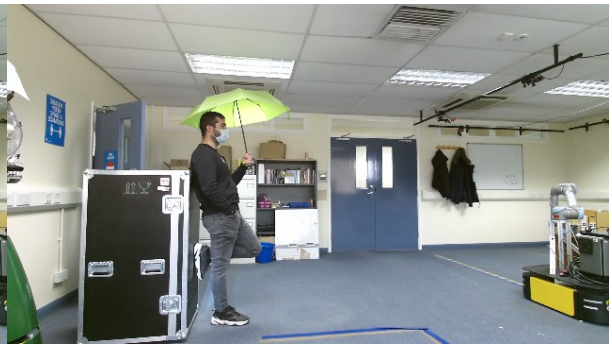


chair: supportable, pullable



suitcase: holdable, carriable, can-cover

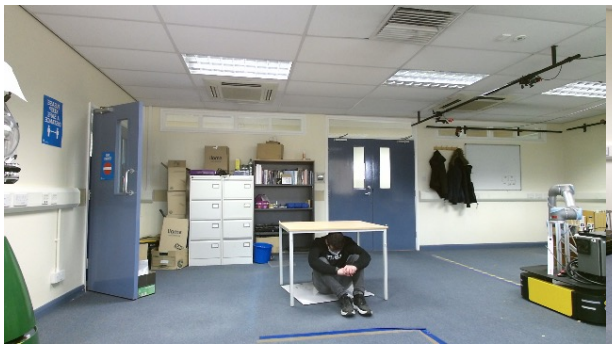
umbrella: holdable, can-cover



box: supportable, can-support, coverable, sittable

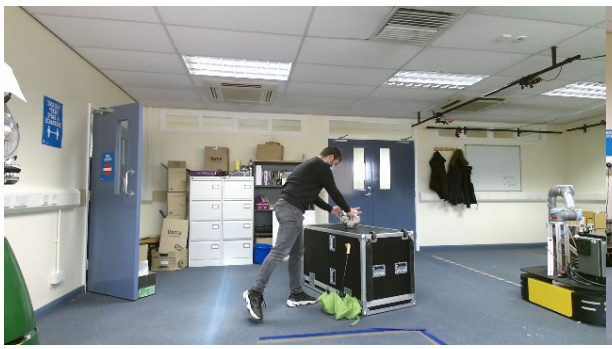
chair: supportable, sittable, coverable

table: supportable, can-cover

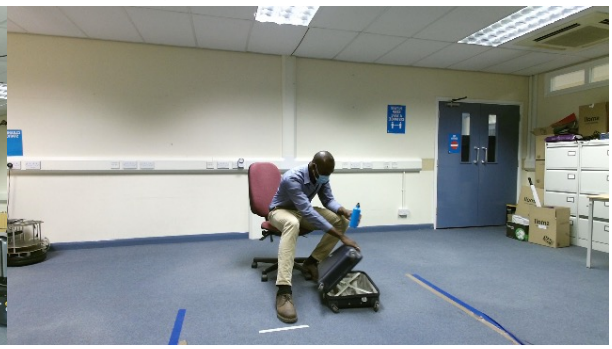


bottle: holdable, drinkable

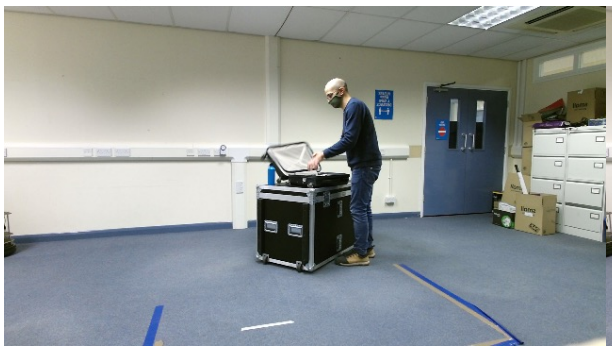
box: supportable, can-support, leanable



suitcase: supportable, can-contain



bottle: supportable, containable, holdable



suitcase: supportable, can-support, coverable, rollable

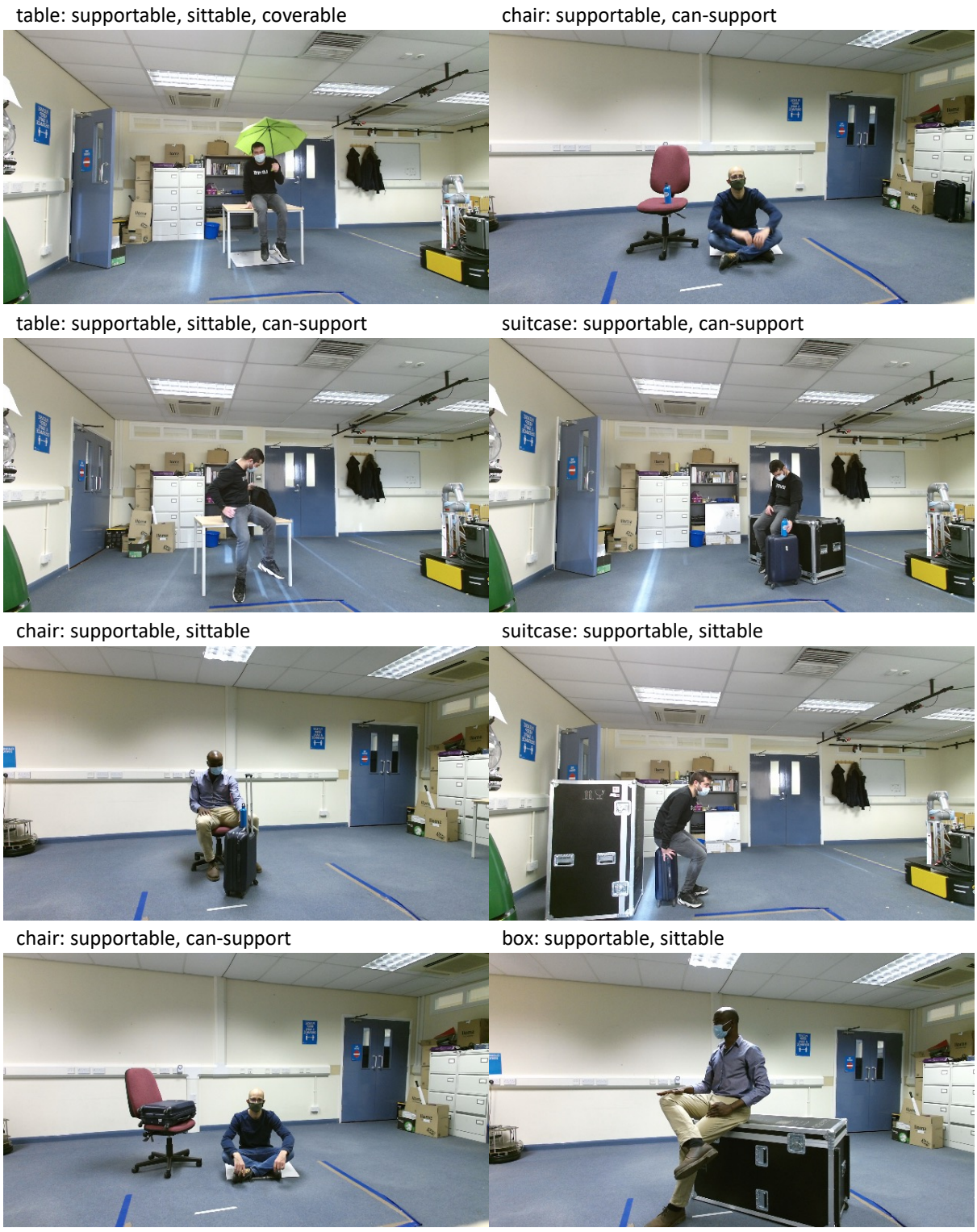


Figure 4.1: LOAD activity samples with affordance labels for one of the objects in the scene.

Figure 4.2: Activities captured in LOAD.

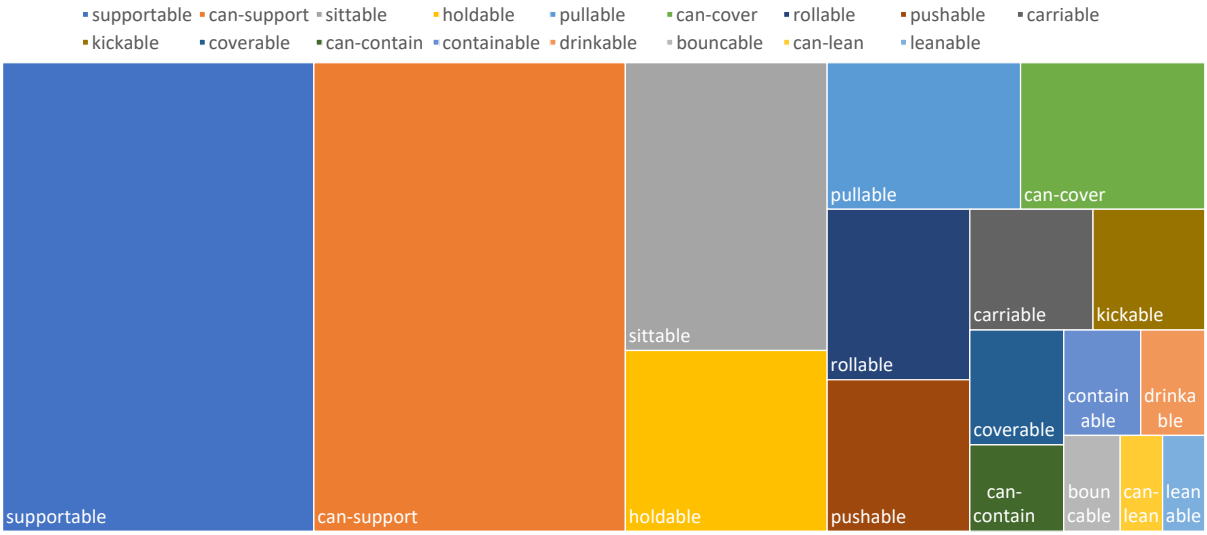


Figure 4.3: Affordance labels in the LOAD dataset; area of rectangle reflects abundance in the dataset.

Moreover, the main benefit of LOAD is the variety of *unconventional* object interactions captured. *I.e.*, the interactions between objects illustrated in the dataset are not limited to the objects' functionalities, but considering their morphological characteristics the objects are utilized by the human agents in various ways, *e.g.* using a suitcase as a cover though its main functionality is to contain. Hence, the same affordance label is present in different human activities, as well as between different kind of interacting objects. Therefore, allowing the differentiation of affordances based on the objects' interactions, rather than the recognized activity or the detected objects in the scene. Furthermore, having the same affordance in various activities prevents overfitting the affordance data on the objects and the scenes, and hence aims at the creation of an affordance space, which is not object-, activity-, or scene-constrained.

Figure 4.4 visualizes the diversity of affordance labels in every activity present in the dataset. The figure also displays the percentage of occurrence of the affordances in every activity kind. For every activity in every video of the dataset, affordances are considered for all pairs of interacting objects involved in the specific activity, and only one affordance can hold between a pair of object at a time. *I.e.*, during a "drink" activity, since the involved objects will be used

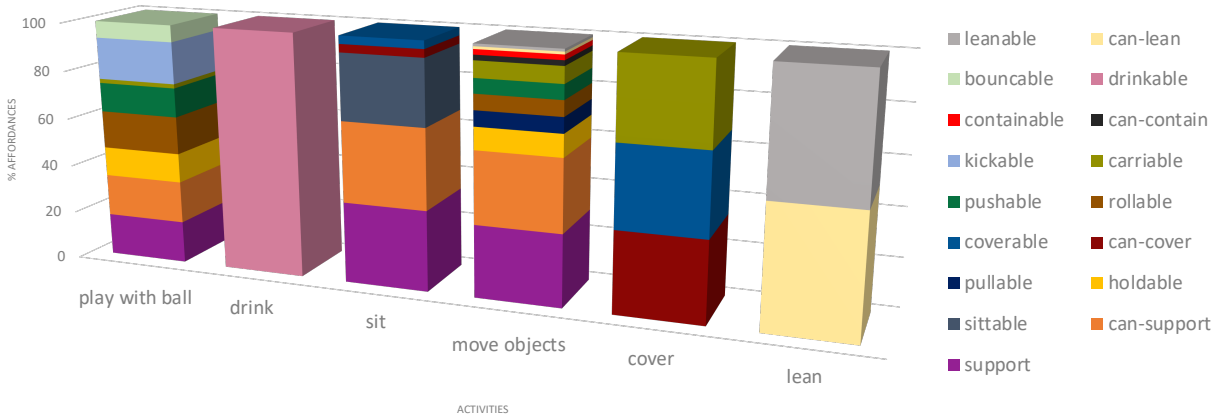


Figure 4.4: Analysis of the percentage of affordances for every activity in LOAD.

for drinking, only the "drinkable" affordance can hold. Some of the affordances that are present in multiple activities are "support", "can-support", "holdable", "rollable", "coverable", "can-cover", whereas affordances such as "can-lean", "leanable", and "drinkable" are only present in the activities "lean" and "drink", respectively.

4.4 Conclusions

The LOAD dataset extends the everyday-life activities domain with video data capturing complex real-world everyday-life activities. This dataset was recorded to enhance existing benchmark video datasets and can be used for any task that relies on human-object interactions, *e.g.* object affordance prediction, *etc.*

Chapter 5

Object Affordance Categorization

5.1 Introduction

Object *affordances* were first formally defined by Gibson (Gibson 1977), stating "the affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill. (...)I mean by it something that refers to both the environment and the animal in a way that no existing term does. It implies the complementarity of the animal and the environment". However the concept of object affordances as it is understood in Computer Science has not been explicitly defined in the literature, causing some confusion. 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 an object, which can be afforded in a specific way (Nguyen et al. 2017; Do et al. 2018; Sawatzky et al. 2017; Kokic et al. 2017; Nguyen et al. 2016; Myers et al. 2015; R. Xu et al. 2021), *e.g.* the handle of a hammer has the affordance of "hold" whereas the head has the affordance of "hit", the inner surface of a cup has the affordance of "contain" although the cup's handle has the affordance of "hold". Such definition, does not allow generalization to an open-set of objects. *E.g.* a book does not have a handle part, however the "hold" affordance can hold. In contrast, in human-object interaction recognition tasks, *affordance* is captured by the way an object is utilized by the human in a scene (Gkioxari et al. 2018; Yao et al. 2013; Fang et al. 2018; Kjellström et al. 2011; Hou et al. 2021; H. Wu et al. 2020; Tan et al. 2019; Qi, Huang, et al. 2017; Chuang et al. 2018), *e.g.* if a human uses a cup for containing something then the cup will have the affordance of "contain". Moreover, an 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. Such multi-labeled affordance objects can be recognized by considering their interactions with other objects.

Hence, the following definition of *affordance* is proposed:

Definition 1

An affordance is a property of an object arising from its interaction with another entity, *i.e.* agent, object, capturing how the object is being used.

Thus, an affordance is correlated with the occurring interaction as every interaction exploits at least one object affordance, *i.e.* a role of the object. Also, an object may have multiple affordances based on the various interactions it may have with other entities. *E.g.* a microwave may have the role of a container if an object is detected to be contained in it, as well as it may act as a supporter if an object is placed and left ontop of it.

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 scenes comprise an open-set of 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 with 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 task, thus changing its primary affordance.

Based on the proposed definition, the problem of affordance inference is addressed, by exploiting object interactions through RGB-D video data. Hence, the detection of affordances relies on the way objects are being utilized by the human agents in a scene, by considering their interactions with other objects as well as the human agent, allowing objects to support different kinds of affordances at the same time.

This chapter is organized as follows. First, a literature review on the task of detection and categorization of object affordances is presented in Section 5.2. The use high-level information for describing object interactions, which relate to their affordances, is shown in Section 5.3, and learning a hierarchy of categories of object affordances in an unsupervised way, by ex-

ploiting graph structures of object interactions, is demonstrated in Section 5.4. Moreover, the experiments conducted to perform evaluation of the proposed approach on various datasets and against several baselines, are presented in Section 5.5, and some discussion and qualitative results are provided in Section 5.6. Finally, Section 5.7 comprise the conclusions of this chapter.

5.1.1 Overview

Figure 5.1 illustrates an overview of the method presented in this Chapter. The proposed method addresses the task of object affordance categorization based on the observed interactions between objects. It focuses on learning a hierarchy of groups of high-level object interactions, by taking into account their spatio-temporal relations from extracted visual appearances. Graphs are able to represent both static and dynamic high level relations between objects. Accordingly, these object interactions are embodied through a high-level graphical structure, the *Activity Graph*, abstracting from the continuous spatio-temporal representation and acquiring depth-informed qualitative spatial relations between object pairs. With these high-level qualitative relations, the graphical structures for capturing interactions are not dependant on scene-specific characteristics, *i.e.* the objects' labels, distance between objects, *etc.*

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 graph similarity using the cosine similarity measure of their embeddings in a learned latent space. By clustering graph structures, a hierarchical tree representation is produced demonstrating their similarity. Since the proposed 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, the depth information is exploited 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 possible ways they can interact with other objects, *e.g.* non-concave objects can not act as "containers".

Hence, the objective of this work is to introduce a novel unsupervised affordance categorization framework, which handles an open-set of interactions and affordances, considering high-level information of the human-object and object-object interactions.

Figure 5.1: Overview of the proposed approach for open-set object affordance categorization, using unsupervised learning, by exploiting high-level object interactions. A sequence of RGB-D image frames is employed as an input to the method. Object proposals, *i.e.* objects' bounding boxes, are extracted from the RGB video frame data using an object detector. The corresponding depth information for every detected objects is exploited to infer the convexity type of every object. The QSR Library uses the detected object proposals with their convexity type to construct *AGraphlets*, which capture a proposed set of spatio-temporal relations. The Graph2Vec network is trained to project these graphs on a latent space. Graph embeddings of *AGraphlets* are then hierarchically clustered to create groups of objects with similar affordances.

5.2 Literature Review

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 and geometric features (Deng et al. 2021; A. Y. Wang and Tarr 2020). 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 (Montesano and Lopes 2009). New approaches employ Convolutional Neural Network (CNN) models to produce classes of functional object parts from RGB (Nguyen et al. 2017; Do et al. 2018; Sawatzky et al. 2017) and synthetic data (Kokic et al. 2017). However, depth cues along with the RGB information have demonstrated a greater detection accuracy in this task (Nguyen et al. 2016; Myers et al. 2015; R. Xu et al. 2021). Additionally, incorporating knowledge about the scene and context in which an object is being used boosts the prediction accuracy even more (Yibiao Zhao and S.-C. Zhu 2013).

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 (Gkioxari et al. 2018; Yao et al. 2013; Fang et al. 2018; Kjellström et al. 2011; Hou et al. 2021; H. Wu et al. 2020). Depending on the human-object interaction being held, a different affordance is detected. Also, prediction of object affordances, cast as human-object interactions, utilize graph neural networks, which are trained on video data to create Object Affordance Graphs (OAG) (Tan et al. 2019). This specific kind of graphs are an intermediate representation and capture the spatio-temporal relations between human and objects/scenes across the input video data through an iterative process. Similarly, a spatio-temporal And-Or graph (ST-AOG) exploits the context captured by objects, actions, and affordances to solve the task of activity understanding (Qi, Huang, et al. 2017). Another approach predicts action-object affordances based on the contextual information of the scene, whilst deploying a graph to represent the objects in the scene as the nodes and their spatial relations as the edges (Chuang et al. 2018). These works demonstrate that by fusing knowledge about the scenario in which an interaction takes place facilitates the prediction of affordances, however limits the generalizability across different domains.

To facilitate domain independence, graph representations of interactions are sometimes em-

	Input data type	Information exploited [a]				Solve task[b]		
		L.F.	H.A.	H.F.	O.T.	O.O.	O.S.A.	O.S.O.
Montesano and Lopes 2009	2D & Keypoints	✓						
Yibiao Zhao and S.-C. Zhu 2013	2D & 3D	✓						
Myers et al. 2015	2.5D	✓				✓		
Nguyen et al. 2016	2.5D	✓						
Nguyen et al. 2017	2D	✓						
Kokic et al. 2017	Synthetic	✓				✓		
Sawatzky et al. 2017	2D & Keypoints	✓						
Do et al. 2018	2D	✓						
A. Y. Wang and Tarr 2020	2D	✓						
Deng et al. 2021	3D	✓				✓		
R. Xu et al. 2021	2.5D & Keypoints	✓				✓		
Turek et al. 2010	2D				✓		✓	✓
Qi, Y. Zhu, et al. 2018	3D			✓				
Kjellström et al. 2011	2D		✓			✓	✓	
Yao et al. 2013	3D		✓					
Qi, Huang, et al. 2017	2.5D		✓	✓				
Gkioxari et al. 2018	2D		✓					
Fang et al. 2018	2D			✓				
Chuang et al. 2018	2D		✓	✓				
Tan et al. 2019	2D		✓	✓				
H. Wu et al. 2020	Synthetic		✓					
Hou et al. 2021	2D		✓					
Sridhar et al. 2008	2D			✓			✓	✓
Aksoy, Abramov, Wörgötter, et al. 2010	2D			✓			✓	
Aksoy, Abramov, Dörr, et al. 2011	2D			✓			✓	
Pieropan et al. 2013	2.5D			✓			✓	
Pieropan et al. 2014	2D			✓	✓			
Moldovan and De Raedt 2014	Synthetic					✓		
Liang, Yibiao Zhao, et al. 2016	2.5D			✓	✓	✓		
Liang, Y. Zhu, et al. 2018	2.5D				✓	✓		
This work	2.5D			✓		✓	✓	✓

[a] L.F.: Low-level features H.A.: Human Actions H.F.: High-level Features (graph representations, embeddings) O.T.: Object tracks

[b] O.O: Object Occlusions O.S.A.: Open-set of affordances O.S.O: Open-set of objects

Table 5.1: Related works on object affordance detection.

ployed. 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 (Qi, Y. Zhu, et al. 2018), whereas outdoors scenes are examined by considering the behavior of moving objects around them (Turek et al. 2010). 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-labeled object affordances.

Similar to the proposed approach, the work of Aksoy, Abramov, Wörgötter, et al. 2010, focuses on the employment of high-level graphs to describe interactions between objects. These graphs represent the structural changes happening in the scene due to the objects' interactions. Though graphical structures are able to extract high-level information about the interactions in the scene, their structure might be the cause of domain restriction (Aksoy, Abramov, Wörgötter, et al. 2010; Aksoy, Abramov, Dörr, et al. 2011; Pieropan et al. 2014; Pieropan et al. 2013). For this purpose, qualitative spatio-temporal relations are exploited for extracting affordances from learned activity event clusters. Closely related to the proposed method is the work of Sridhar et al. 2008, which exploits qualitative spatio-temporal relations in a graph representation of sequential object interactions to categorize objects based on their interactions. 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 (Liang, Y. Zhu, et al. 2018; Liang, Yibiao Zhao, et al. 2016; Moldovan and De Raedt 2014). However, these approaches are restricted to the detection of a single object affordance, *i.e.* containment, and prone to false positive detections of the containment relation due to occlusions.

Table 5.1 presents the related works of object affordance detection and prediction, summarizing the kind of information each works employs and the task it is focusing on.

5.3 Representation of Object Interactions

5.3.1 Qualitative Object Interaction Graphs

Relational graph structures represent high-level information by abstracting from the continuous space of the exploited relations. 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 and human-object interaction reveals each object's affordances. Relational graph

RCC2	Definition	Description
$\mathsf{C}(\alpha,\beta)$ ‡	$mask(\alpha) \cap mask(\beta) \neq \varnothing$ †	α and β are connected
$\mathsf{DC}(\alpha,\beta)$	$\neg\mathsf{C}(\alpha,\beta)$	α is disconnected from β

† $mask(x)$ defines object x's region in the image plane, that represents a collection of pixels.
‡ If either α or β comprise a single pixel, then C holds if a pixel from α and a pixel from β are adjacent. Whereas, if α and β comprise of more than 2 pixels, then C holds when there is at least one pixel shared between them.

Table 5.2: RCC2 relations.

structures of object interactions aid in representing an open set of interactions. Thus, the representation of *Activity Graphs* (*AG*s), presented in Section 2.3, is exploited.

In this work, *AG*s are deployed to capture object-object and human-object interactions, thus two different qualitative spatial relations (QSRs) are considered, one for each kind of interaction. A set of novel QSRs, the DiSR set introduced in Chapter 3, is employed to describe the spatial relationships between objects, as it can effectively capture complex object interactions such as, "contain" and "support". Whereas the RCC2 set of relations (D. A. Randell et al. 1992; Anthony G Cohn et al. 1997) (Table 5.2) are used for representing interactions between objects and humans, to incorporate the information of a human initiating an interaction by grasping an object.

These two spatial relational sets comprise the V_{spat} layer of the *AG*s, whilst *Allen's temporal algebra* (Section 2.2.3) is exploited to express the temporal relationships between the spatial relations (V_{temp}).

An *Activity Graphlet* (*AGraphlet*) of an object is defined as a sub-graph of an *AG*, which carries the spatial and temporal information ($V^{'}_{spat}$, $V^{'}_{temp}$) of the respective object's interactions with another object and a human body part ($V^{'}_{ent}$). Figure 5.2 illustrates how *AGraphlet*s are extracted from a video sequence capturing interacting objects. Firstly, qualitative spatial relations are extracted from detected episodes in the temporal domain of a video. *AGraphlet*s are constructed using these qualitative spatio-temporal relations for individual detected objects in the scene describing their interaction with another object and a human body part, *e.g.* a human hand.

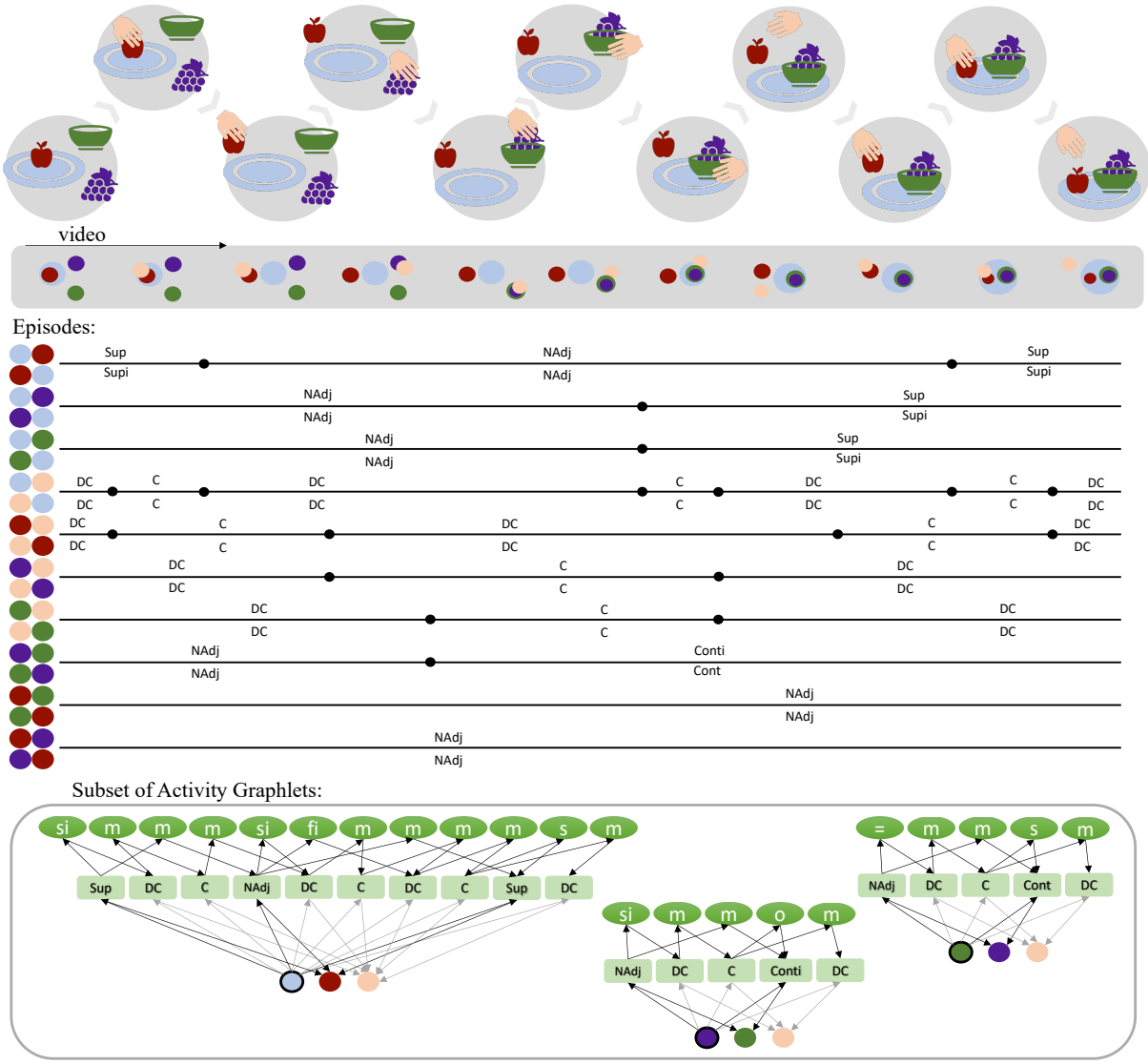


Figure 5.2: (best viewed in color) *AGraphlet*s are extracted from a video sequence of interacting objects. Firstly, qualitative spatial relations are captured from the detected episodes in the temporal domain of a video. *AGraphlet*s are constructed using these qualitative spatio-temporal relations for individual detected objects (encircled) in the scene describing their interaction with another object and a human body part, *e.g.* human hand. For simplicity, only a subset of the whole set of *AGraphlet*s is visualized in this figure.

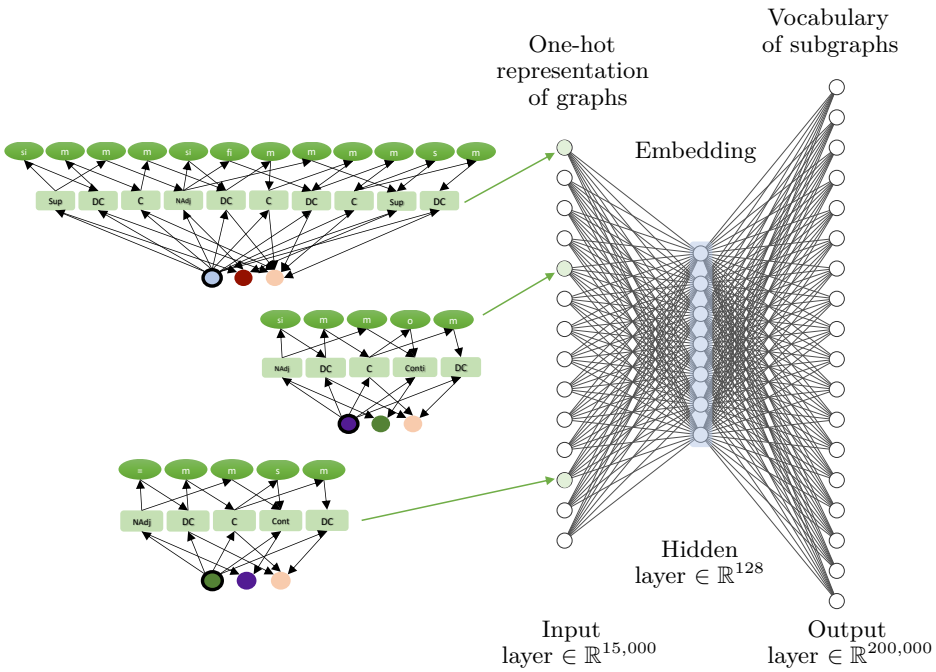


Figure 5.3: *Graph2vec* network is employed to project the extracted *AGraphlet*s, represented as a one-hot representation, on a learned latent space. The size of the depicted network is specific to one of the folds used for training.

5.3.2 Graph Embeddings of Object Interactions

Graph embeddings are d-dimensional vector representations of projections of graphs in a d-dimensional latent space[1]. By projecting graphs in a multi-dimensional space one is able to find similarities/dissimilarities between them by considering distance measures on their vector representations.

Inspired by the Natural Language Processing literature, the *graph2vec* network (Narayanan et al. 2017) is exploited, to learn graph embeddings in an unsupervised way and project the *AGraphlet*s in the learned latent space. The network employs the notion of context, where context is defined by a fixed number of subgraphs comprising every *AGraphlet*. *AGraphlet*s with similar subgraphs represent similar affordances. Figure 5.3 [2] shows an illustration of the *graph2vec* network with the *AGraphlet*s as input data. During training time, each *AGraphlet* corresponding to a unique input node, is mapped to the appropriate subgraphs from the vocabulary of subgraphs in the output space. Thus, at inference time a one hot representation of the input graph, activates the node with which it most similar to.

[1] A latent space is a high-dimensional abstract space that encodes an internal representation of observed data. Thus, similar data points are positioned close in the latent space.

[2] The size of the depicted network is specific to one of the folds used for training.

Fine-tuning *graph2vec* Network

The *graph2vec* network was trained with the *AGraphlet*s of the training set to create a 128-dimensional latent space which the interaction graphs can be projected onto. The network architecture employed was provided by Narayanan et al. 2017 and was trained using Stochastic Gradient Descent, a batch size of 512, and setting the learning rate to 0.5. This shallow encoder-decoder network exploits the Weisfeiler-Leman kernel (Shervashidze et al. 2011) to extract sub-tree patterns of the input graphs, as it has been demonstrated that it outperforms linear substructure-based kernels, *e.g.* random walk, shortest path kernel (Yanardag and Vishwanathan 2015; Shervashidze et al. 2011). The Weisfeiler-Leman subgraph extraction value was set to 14 through an empirical study, and it was found to be sufficient to create all the different subgraphs from the *AGraphlet*s.

For the selection of the embedding size, the *graph2vec* network was trained with the embedding size values 128, 256, 512, and it was concluded that an embedding of size 128 produces the smallest training and validation loss. Figure 5.4 illustrates the training and validation loss of *graph2vec* for the different embedding sizes. The loss history does not update if a validation loss value greater than the latest reported has been calculated. Batch size (512) and learning rate (0.5) remained unchanged in this experiment.

Moreover, an evaluation was conducted of the output after training the network with different learning rates (0.3, 0.5, 0.7). As illustrated in Figure 5.5, the best training results were produced with learning rate 0.5. The batch size (512) and embedding size (128) remained the same throughout this study. The batch size was selected through an empirical study.

5.4 Unsupervised Learning of *AGraphlet*s

To learn a hierarchy of object affordances a hierarchical clustering approach is considered, acting on the *AGraphlet*s. Every *AGraphlet* represents the interaction between a pair of objects in the scene. An interaction between two objects is considered as the spatio-temporal sequence of relations holding during an activity. A hierarchy of groups of similar affordances is produced by clustering such graph structures, which are closely related with the way every object is being used in an activity. Hence, the proposed method does not pose any constraints on the number of affordance clusters an object can be assigned to, whilst every object has as many *AGraphlet*s

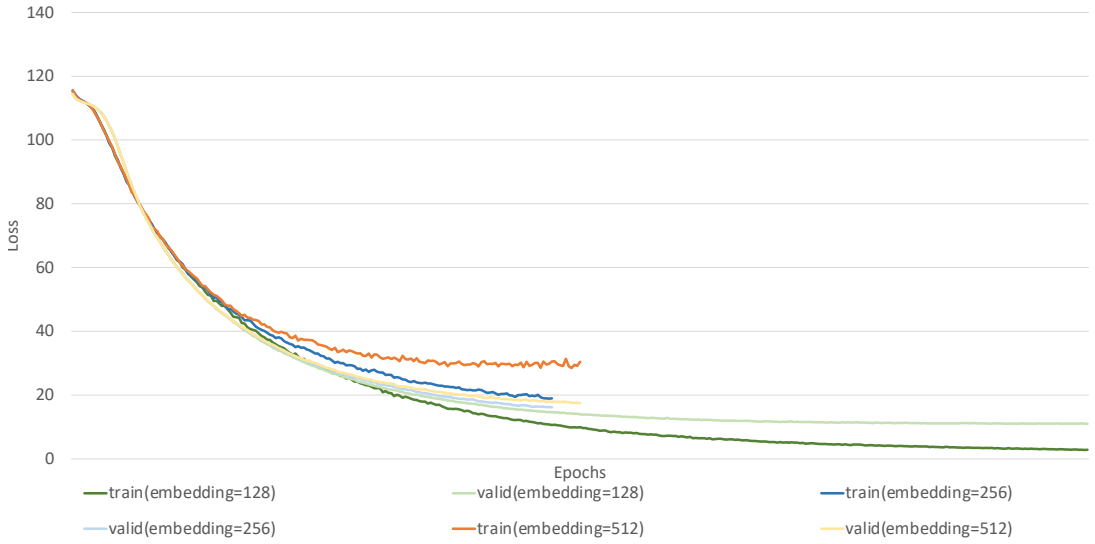


Figure 5.4: (best viewed in color) Training and validation loss of the *graph2vec* network for different embedding sizes. The experiments were done with batch size 512 and learning rate 0.5. The loss history does not update if a validation loss value greater than the latest reported has been calculated.

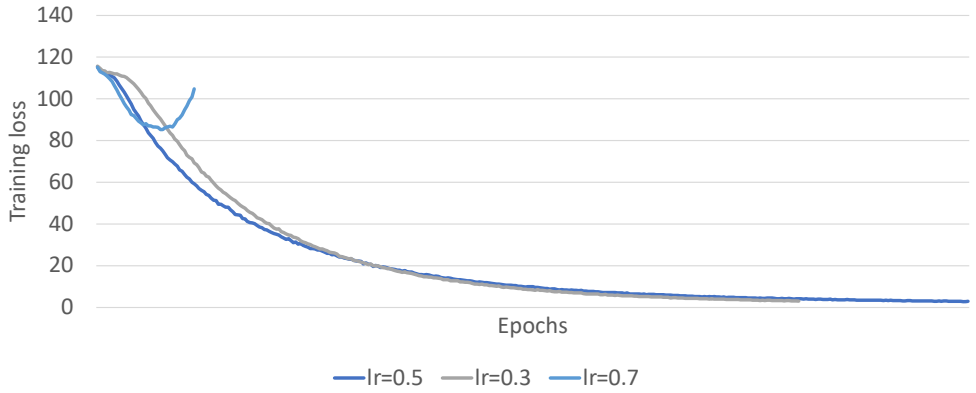


Figure 5.5: (best viewed in color) Training loss of the *graph2vec* network with embedding size 128 for different learning rates. The experiments were done with batch size 512. The training loss history does not update if a validation loss value greater than the latest reported has been calculated.

as detected interactions. *E.g.* consider the scenario where an agent picks a bowl from a table and places it in the microwave. The proposed approach considers the interaction of the bowl with the table and the bowl with the microwave as two different interactions; two *AGraphlet*s will be created for the bowl implying that it is a "supportable" as well as a "containable" object.

As *AGraphlet*s are high-level object interaction representations defined in the space of the qualitative relations being captured, computing the similarity between them is based on the spatio-temporal relations occurring. To effectively cluster these representations the *AGraphlet*s are projected in a d-dimensional latent space and clustering is performed on that space instead. The *graph2vec* network is employed for this purpose, creating d-dimensional graph embeddings for each *AGraphlet*.

To compute the similarity/dissimilarity of these graph embeddings a cost function is exploited, using the cosine similarity measure, and it is defined as:

$$\mathfrak{L}(\mathrm{A},\mathrm{B}) = 1 - cos(\mathrm{A},\mathrm{B}) = 1 - \frac{\mathrm{A}\cdot\mathrm{B}}{||\mathrm{A}|| \times ||\mathrm{B}||} \tag{5.1}$$

where A and B represent the embeddings of two *AGraphlet*s, and the values of the $cos(\cdot)$ function range from -1 to 1. 1 indicates that the vector embeddings are exactly the same, 0 that they are orthogonal, and -1 means that the vectors are exactly opposite. In the *AGraphlet* space, a cosine value of 1 stipulates that the two graphs carry the same spatio-temporal information, hence the same affordance, whereas a value of -1 suggests that the two graphs have no spatio-temporal information in common, thus each one of them represents a completely different affordance. A cosine value of 0 denotes that the two graphs are opposite, each one carrying information of the two objects that interact. A hierarchical agglomerative[3] clustering on the graph embeddings is performed, using the *cosine similarity* as the distance measure, to produce a dendrogram of similarities of the extracted *AGraphlet*s. Clusters are formed by grouping the leaves of the dendrogram with respect to their hierarchy. Figure 5.8(top) illustrates a subset dendrogram of the complete hierarchical clustering output. Such a hierarchy reveals the similarities of the different interactions occurring in the data, which thus yields a set of affordance clusters. Different set of clusters are formed based on a tunable threshold in the hierarchy. The higher in the hierarchy the creation of clusters is considered, the more complete, however less homogeneous the clusters

[3]Agglomerative clustering is a bottom-up approach, *i.e.* initially each observed data point is its own cluster and iteratively clusters are merged whilst moving up-wards in the hierarchy.

become, in terms of affordances.

5.5 Experimental Evaluation

5.5.1 Experimental Setup

Datasets

For the evaluation of the proposed method the publicly available CAD-120 dataset (Koppula et al. 2013) (CAD) is used. Moreover, to prove generalizability across different data distributions the proposed approach is also tested on the Watch-n-Patch (C. Wu et al. 2015) (WnP), and on the LOAD dataset (LOAD). All three datasets comprise activities of everyday-life scenarios with various configurations of the objects in the scene as well as different camera orientations. Chapter 4 provides more information about the data of these datasets.

80% of the CAD-120 dataset is used to determine the method's parameters and train the *graph2vec* network, resulting in 18,072 *AGraphlet*s, and the proposed approach is evaluated on the remaining 20% unseen videos (24 videos), which comprise of 5,682 *AGraphlet*s. The dataset split was performed after random shuffling the video data of all the human subjects and the activities performed in the dataset. Experiments on the Watch-n-Patch and LOAD datasets are conducted on 24 and 15 hand-picked videos, or 6,900 and 1,212 *AGraphlet*s respectively, where the predictions of the object detector, after visual inspection, are sufficient for capturing interacting objects. For defining the $thresh_{convex}$, sd, and sd_{max} parameters in Algorithm 1 and Algorithm 2, for the detection of the DiSR relations, hand-picked videos are exploited from the training set for the CAD-120 dataset, and for the LOAD and Watch-n-Patch datasets hand-picked videos are used different from the ones used for testing the proposed method.

Human/Object Detection & Tracking

To obtain human skeletal data in the scene a comparison was done between the output of the Convolutional Pose Machine (CPM) (Wei et al. 2016) model and the Kinect skeletal data. Through an empirical study it was found that the CPM predictions are more accurate, especially when human-joint occlusion occurs or the human agent stands sideways.

Object locations and depth information are provided from the predicted objects' masks. The state-of-the-art two-branch Mask R-CNN framework (He et al. 2017) was employed, which was

trained on the COCO dataset (Lin et al. 2014), since the COCO training data distribution is similar to the data distribution present in the datasets used in this work, and it is a generic dataset with more object classes than those included in the target task. The box enclosing the object's mask corresponds to the object's bounding box. This implementation is based on Mask R-CNN predictions, however the proposed method is not object-specific and any class-agnostic proposal method can be used.

Whilst Mask R-CNN objects' bounding boxes predictions are temporally sparse, the objects tracks are enriched by considering the CSRT tracker's predictions (Bolme et al. 2010) from the latest object bounding box occurrence, for frames where Mask R-CNN failed to detect the object, hence creating tracks of class agnostic-masks. The CSRT tracker solely relies its predictions of the visual characteristics of the initial object mask predictions, considering only past instances on mask detections. Hence, it is able to track non-rectangular objects, which aids in having more accurate tracking predictions by considering the objects mask's rather than their bounding boxes. 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.

The predicted object mask's depth information is employed to infer the object's convexity type for the DiSR relation detection. Any pixel, which is part of a human detection, is excluded by retrieving a human mask from the DensePose framework (Alp Güler et al. 2018). 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.

Relations & Clustering Parameters

The QSRlib library (Gatsoulis et al. 2016) is employed for the construction of *AGraphlet*s. For the sd and sd_{max} values which are used in defining the *indentation area* of a concave object, the values of 5 and 3 are selected respectively, after conducting an empirical study for $sd \in \{2, 3, 4, 5, 6, 7\}$ and $sd_{max} \in \{1, \ldots, h-1\}$. Also the $thresh_{convex}$ value is set to be 4 for the CAD-120 dataset after evaluating on the values 1,2,3,4,5,6,7,8,9,10, 0.3 for the Watch-n-Patch

after consideration of the values in the range 0.1 to 2.0 with step 0.1, and 10.0 for the LOAD after studying the values in the range 8 to 20 with step 1.

For the creation of clusters the produced hierarchy is thresholded at 0.02 for all three datasets. This hierarchy threshold is defined by evaluating the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) on the training set for different heights of the hierarchy. The BIC and AIC criterions are defined as:

$$BIC = \theta \cdot ln(n) + 2 \cdot ln(\hat{L}) \tag{5.2}$$

$$AIC = 2 \cdot \theta - 2 \cdot ln(\hat{L}) \tag{5.3}$$

where θ is the model's parameters, n is the number of data points, and $\hat{L}$ represents the likelihood function. The likelihood function is the probability that the data are explained by the model. In this work the clustered data are discrete data, hence the probability is correlated to the entropy of the data. The *V-measure* score function, provides a measure, ranging from 0 to 1, of how well the data are partitioned by considering the entropy of partition of clusters. Thus, in this case the *V-measure* score function is used as the likelihood function.

5.5.2 Quantitative Evaluation

This section provides an analysis of the experimental results by inspecting the clusters of affordances formed, and evaluating their homogeneity and completeness. The data points clustered consist of detected objects with their interactions with any other object and the human agent, allowing multiple data points to point to the different interactions a single object might hold. The evaluation of the clusters comprise of the reporting of the normalized *V-measure*, *homogeneity*, and *completeness* scores, with higher values implying a better clustering. Homogeneity captures how homogeneous the clusters are with respect to the ground truth annotations, whereas completeness expresses how complete the clusters are in terms of the ground truth affordance labels, *i.e.* encapsulating as many as possible of the same ground truth labels into a single cluster. The V-measure represents an average of the homogeneity and completeness scores. Section 2.4 provides more details on the calculation of these metrics.

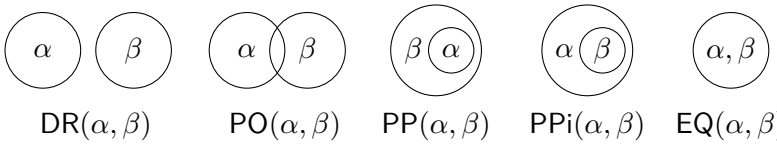
	Method	V-measure	Homogeneity	Completeness
CAD	RCC5(+ON)	0.80	0.93	0.70
	sED	0.57	0.67	0.49
	Proposed	**0.87**	**0.99**	**0.77**
WnP	RCC5(+ON)	0.49	0.83	0.34
	sED	0.14	0.13	0.16
	Proposed	**0.57**	**0.99**	**0.40**
LOAD	RCC5(+ON)	0.62	0.78	0.52
	sED	0.40	0.44	0.37
	Proposed	**0.70**	**0.99**	**0.54**

Table 5.3: Ablation study experiments.

Ablation Study

An ablation study is conducted to evaluate each component of the proposed method. This study evaluates the impact of using depth information to enhance 2D qualitative spatio-temporal relations, and projecting *AGraphlet*s into a latent space. The experimental setups for comparison are:

- RCC5(+ON): evaluate the proposed method with RCC5 (D. A. Randell et al. 1992), which consists of the relations: "discrete" (DR), "partially overlapping" (PO), "proper part" (PP, PPi), and "equal" (EQ), along with the addition of the On spatial relation defined in Chapter 3.3.1. A schematic visualization of the RCC5 qualitative relations is presented here:



  The aim of this investigation is to evaluate how the lack of depth information affects the categorization of object affordances. For this study the method's configuration is kept unchanged and the conducted experiments alter only the exploited spatial relations. For the creation of clusters the produced hierarchy is thresholded at 0.03 based on the evaluation of the BIC and AIC on the training set for different heights of the hierarchy produced;

- sED: perform hierarchical clustering on the *AGraphlet*s without projecting them in a latent space learned by the *graph2vec* network. This setup evaluates the clustering performance

by considering a vanilla measure for graph comparison and aims to showcase the contribution of projecting *AGraphlet*s in a latent space. Hence, a *set Edit Distance* measure is exploited to create clusters of similar affordances. This measure is defined by Equation 5.4, where c_{spat} and k_{spat} are 0.5 from the ablation study presented in Figure 5.6. In Figure 5.6 the spatial weights extend from 0.0 to 1.0, and the plots visualize the homogeneity, completeness, and v-measure scores in the y-axis, for various dendrogram thresholds in the x-axis, *i.e.* how low or high in the dendrogram hierarchy one starts to consider the creation of clusters. The dendrogram thresholds start from 0.0 and finish at the highest similarity score. For the creation of clusters the produced hierarchy is thresholded at 1.0 based on the evaluation of the BIC and AIC on the training set for different heights of the hierarchy produced.

$$\begin{aligned} sED = c_{spat}\sum\nolimits_{v\in V'^{\alpha,\beta}_{DiSR}} v + c_{temp}\sum\nolimits_{v\in V'^{\alpha,\beta}_{DiSR_{temp}}} v + \\ k_{spat}\sum\nolimits_{v\in V'^{\alpha,\beta}_{RCC2}} v + k_{temp}\sum\nolimits_{v\in V'^{\alpha,\beta}_{RCC2_{temp}}} v \\ \text{where } c_{temp} = 1 - c_{spat} \text{ and } k_{temp} = 1 - k_{spat} \end{aligned} \tag{5.4}$$

$$\begin{aligned} V'^{\alpha,\beta}_{R} = \{v : v \in \{V'^{\alpha}_{R} \setminus V'^{\beta}_{R}\} \cup \{V'^{\beta}_{R} \setminus V'^{\alpha}_{R}\}\} \\ \text{where } R \in \{DiSR, DiSR_{temp}, RCC2, RCC2_{temp}\} \end{aligned} \tag{5.5}$$

Table 5.3 summarizes the results from this study. It is evident that the proposed method outperforms the RCC5(+ON) and sED experimental setup in all reported metrics and datasets. The comparison between the experimental results of the proposed approach and the RCC5(+ON), demonstrate that the employment of depth information for inferring object affordances has a considerable improvement of the V-measure in comparison to the primary RCC5 set. This improvement results from more accurate spatial relationships of interactions, enabling the creation of more accurate graphical structures for describing them. Also, the scores' improvement involving the employment of the *graph2vec* network, in comparison to the sED indicates the enhancement of the clustered data point representation by projecting the *AGraphlet*s into the latent space.

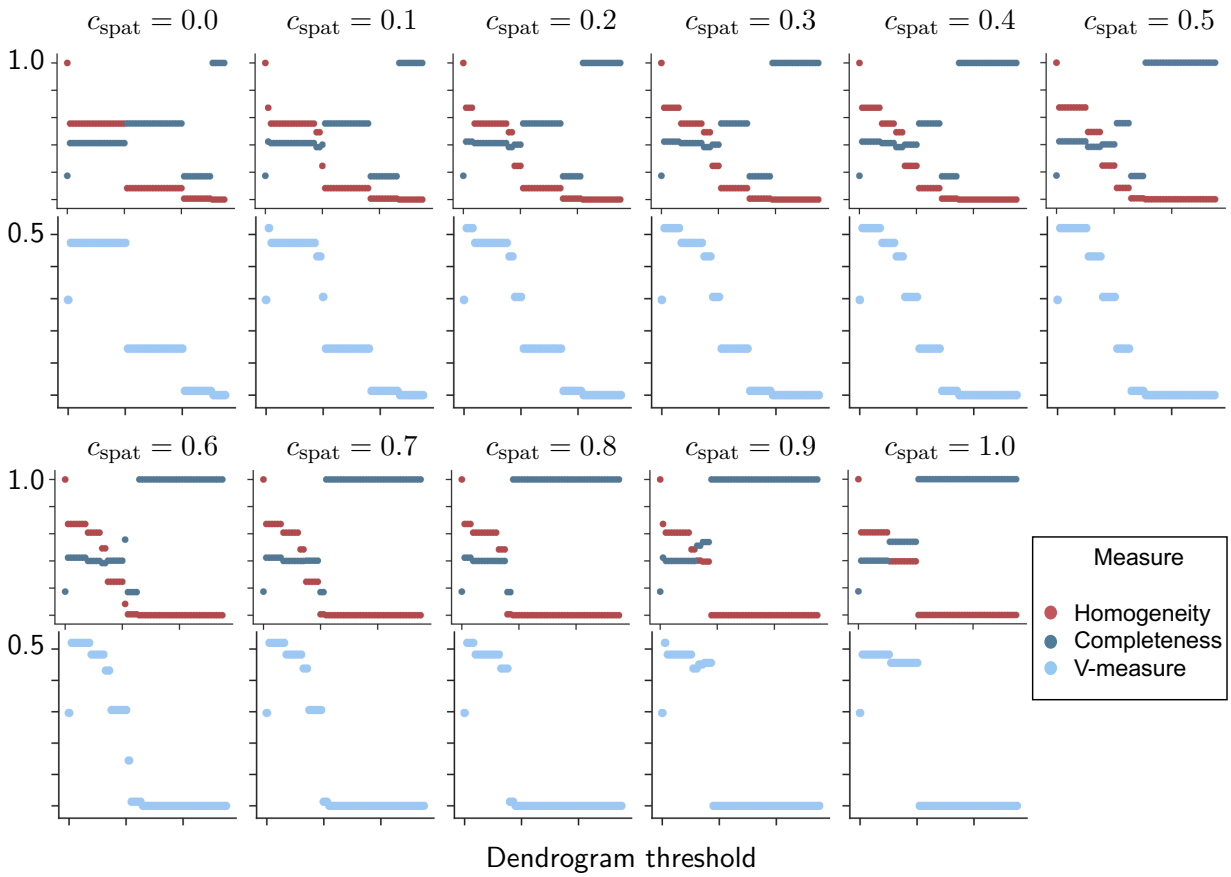


Figure 5.6: Ablation study for defining the sED spatial coefficients.

Baselines Comparison

Related works exploit the distance between objects as a measure for perceiving their interactions (Pieropan et al. 2014), or simplistic qualitative relations, which only describe the relation of the objects with the human hand, *e.g.* "hand approaching", "hand leaving", "in use", "idle", and "close to" (Pieropan et al. 2013). Both approaches are not able to capture complex object interactions. For this reason, the evaluation of the proposed approach considers methods, which leverage qualitative spatial relations for describing human-object and object-object interactions, which are more descriptive in the spatial domain.

The proposed approach is compared with two state-of-the-art methods, in the area of object affordance/functionality categorization and classification based on high-level qualitative representation of object interactions. Specifically, these methods use qualitative spatio-temporal relations to describe, in a high-level way, human-object and object-object interactions, similar to the proposed method.

These works are:

	Method	V-measure	Homogeneity	Completeness
CAD	Sridhar et al. 2008	0.40	0.66	0.29
	Aksoy, Abramov, Wörgötter, et al. 2010	0.71	0.84	0.61
	Proposed	**0.87**	**0.99**	**0.77**
WnP	Sridhar et al. 2008	0.39	0.65	0.28
	Aksoy, Abramov, Wörgötter, et al. 2010	**0.58**	0.70	**0.50**
	Proposed	0.57	**0.99**	0.40
LOAD	Sridhar et al. 2008	0.63	0.82	0.51
	Aksoy, Abramov, Wörgötter, et al. 2010	0.61	0.77	0.50
	Proposed	**0.70**	**0.99**	**0.54**

Table 5.4: Experimental comparison with state-of-the-art works.

- Sridhar et al. 2008: experiments conducted using a re-implementation of the approach of Sridhar et al. 2008 which, to the best of our knowledge, is the most closely related work that exploits QSRs to capture functional object clusters;
- Aksoy, Abramov, Wörgötter, et al. 2010: performed comparison of the proposed method with a re-implementation of the approach of Aksoy, Abramov, Wörgötter, et al. 2010, which exploits high-level qualitative graphs to describe interactions between objects.

Both of these works employ a set of spatial qualitative relations. However, these relations are not descriptive enough to distinguish between complex object interactions, *i.e.* containment and support. Hence, differentiation of affordances due to different object types is not present, resulting to a coarser-grain of affordance categorization. Moreover, occlusion is not handled, thus capturing many false positive object interactions. *E.g.* in the aforementioned related works, the depth information of the detected objects is not considered, hence the overlap of bounding boxes of object proposals in the 2D image plane is identified as an interaction between these object proposals, even in the case where the objects are not interacting but one is in front of the other. However, in the proposed approach, due to employment of the DiSR relations, it differentiates between actual interaction, *i.e. containing*, *supporting*, and *touching*, and non-interaction, *i.e. not adjacent*.

Table 5.4 shows the quantitative results of the experiments conducted to demonstrate the performance of the proposed approach against the Sridhar et al. 2008 and Aksoy, Abramov, Wörgötter, et al. 2010 baselines. The presented experiments demonstrate a significant increase in all the reported metrics against the Sridhar et al. 2008 work, which results from the representation of *AGraphlets* that exploit more information about the interactive objects. The proposed

method also achieves higher scores than the Aksoy, Abramov, Wörgötter, et al. 2010 work in the CAD-120 and LOAD datasets, and obtain comparable V-measure score on the Watch-n-Patch dataset. This is due to the fact that not many interactions are evident in the video scenes of the Watch-n-Patch, and the set of relations being captured are more coarse-grained, resulting to a higher completeness score, however degrading the homogeneity of the clusters.

These results indicate that the proposed method creates more fine-grained groups of affordances than the ground truth annotations, *i.e.* there can be more than one group of data points indicating a "supportable" affordance. Hence, different objects with the same convexity type are distinguishable.

5.5.3 Qualitative Evaluation

A representative instance of the object affordance categorization method is visualized through the output of the hierarchical clustering algorithm in Figure 5.7. Though evaluation was performed on a set of videos, *i.e.* the test set, for simplicity in visualization, a subset of 100 *AGraphlet*s from the test set of the CAD-120 dataset is employed in this figure as an input to the proposed methodology. Groups of similar *AGraphlet* embeddings in the latent space form clusters of object affordances. Clusters are color-coded in the presented hierarchy (Fig. 5.7) with the y-axis corresponding to the distance measure and the x-axis showing the cluster identifier every leaf node is assigned to; the edges of the leaf nodes are colored depending on the cluster the leaf node is assigned to. Moreover, the same color code has been applied on the 3D visualization of the *AGraphlet* embeddings of the clustered data (Fig. 5.8(a)), where a PCA has been applied on the d-dimensional *AGraphlet* embeddings to reduce to 3-dimensional data, *i.e.* components 1, 2, and 3, for visualization purposes. These components represent an abstraction of the d-dimensional embeddings.

The depicted clusters denote the different kind of affordances in the present subset of a video. To facilitate better understanding of the results, by considering the dominant ground truth affordance label in each group of data points, an affordance name is given to every cluster (Fig. 5.8(b)). The cluster identifiers mapped to an affordance label share the same color as in the hierarchy respectively. Some clusters have been assigned the "unknown label", which denotes an affordance that is not present in the ground truth data hence, creating a fine-grained affordance hierarchy. Also, some clusters have two affordance labels, which indicates

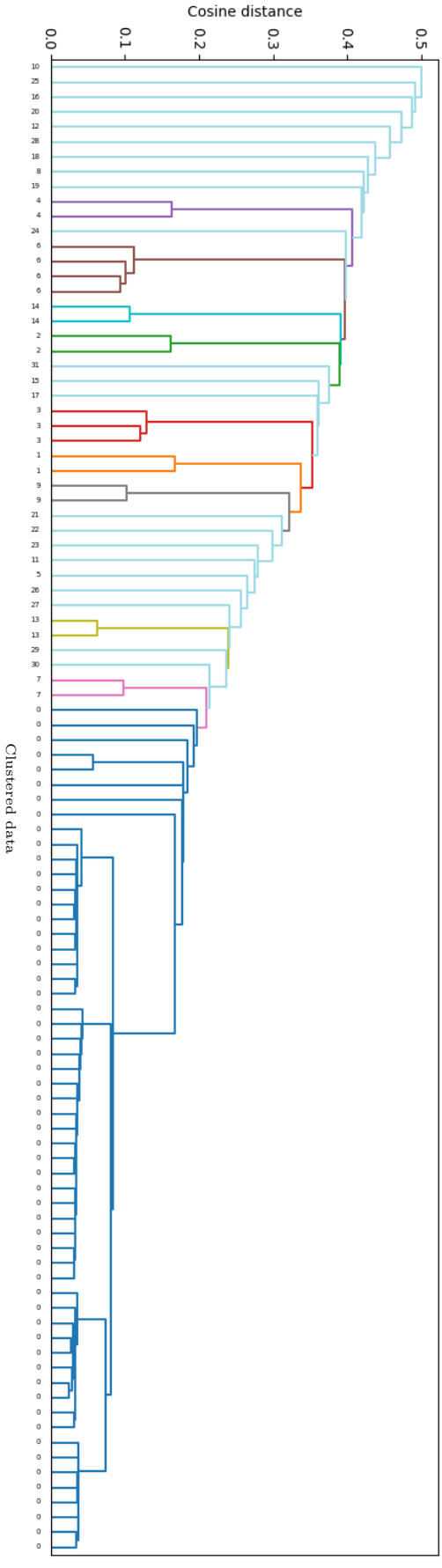
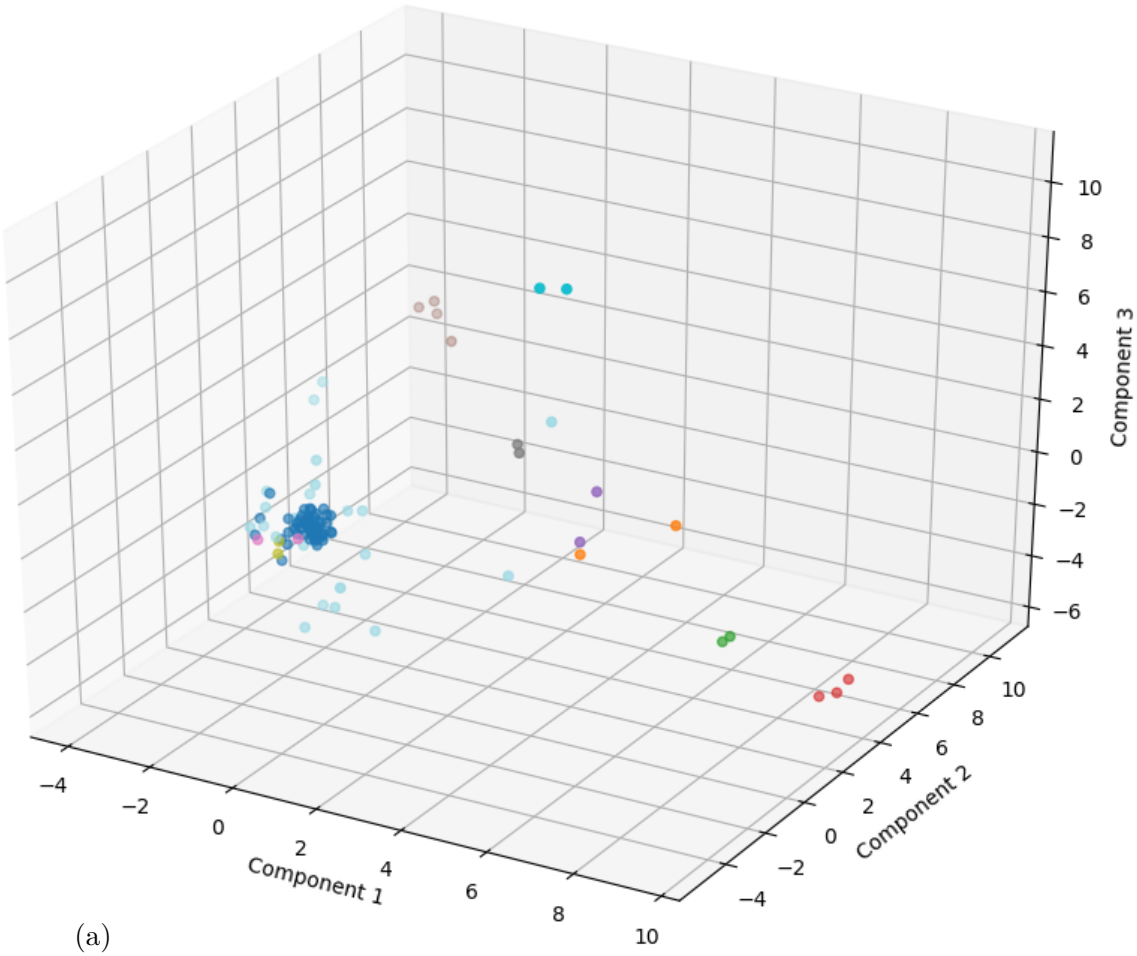


Figure 5.7: Hierarchical clustering of 100 *AGraphlets* captured from the CAD-120 dataset. Dendrogram output from hierarchical clustering where the y-axis corresponds to the distance and the x-axis shows the cluster id every leaf node is assigned to; edges of the leaf nodes are colored depending on the cluster the leaf node is assigned to.



(a)

(b)

Cluster identifier	Ground truth label
0	containable, supportable
1	can-contain
2	can-contain
3	unknown label
4	can-contain
6	can-support
7	supportable
9	can-support
13	can-contain
14	supportable

(c)

V-measure	Homogeneity	Completeness
0.70	0.75	0.66

Figure 5.8: Results of hierarchical clustering of the *AGraphlet*s captured from the CAD-120 video dataset from Figure 5.7. (a) Latent space with color-coded visualizations, based on the leaf color in the output hierarchy (Fig. 5.7), of the graph embeddings' location after a PCA dimensionality reduction. (b) Color-coded cluster identifiers, based on the leaf color from the output hierarchy (Fig. 5.7), mapped to the ground truth affordance labels that best describe the set of *AGraphlet*s enclosing in each group. (c) The quantitative metric scores, *i.e.* V-measure, homogeneity, and completeness, for the specific set of *AGraphlet*s.

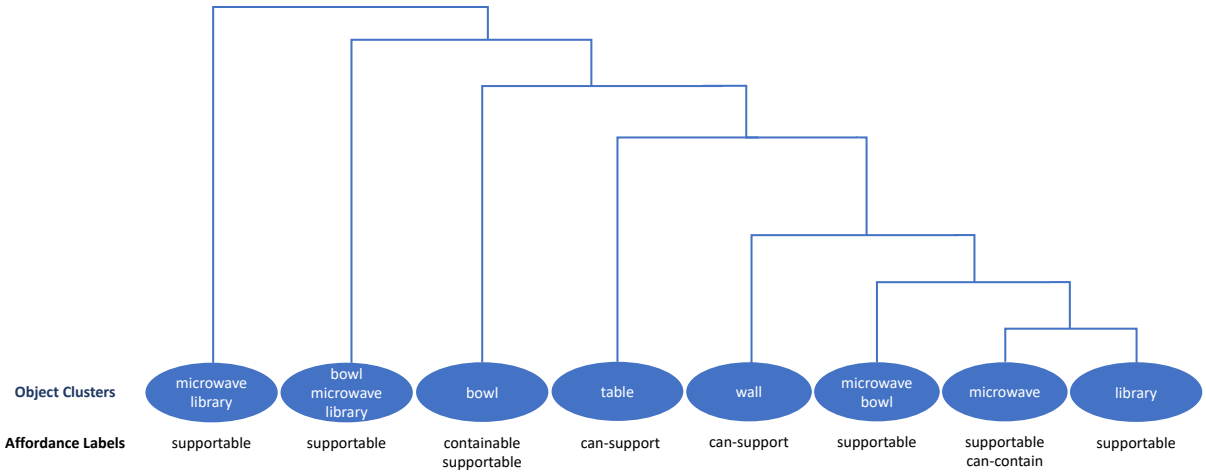


Figure 5.9: Clusters' hierarchy for a video from the CAD-120 dataset.

that the objects in those clusters carry both assigned affordance labels. The produced clustering hierarchy of affordances takes no account of any object labels, thus a cluster is able to contain all objects that have similar interactions in reference to their *AGraphlet* embeddings.

The hierarchical tree presented is generalizable to the whole dataset. More hierarchical trees of other videos demonstrating the generalization of the proposed approach across different datasets, thus videos, are illustrated in Figure 5.10 for the Watch-n-Patch dataset and Figure 5.13 for the LOAD dataset.

Figure 5.9 shows the clustering output of a video from the CAD-120 dataset, with the graph embeddings being replaced by the objects they are referring to. Groups of objects with similar affordances form the clusters of the hierarchy. Ground truth affordance labels are assigned to every cluster by inspecting the objects and their *AGraphlet*s the cluster consist of. The assigned ground truth label for every *AGraphlet* considers the actual affordance the corresponding object has in the real-world, rather than the DiSR relations captured for that specific *AGraphlet*, *i.e.* interaction. These results show the output of a single video for simplicity, however they can generalize across the whole dataset.

An important aspect of the proposed method, as illustrated in this figure, is that a fine-grained affordance categorization is achieved by differentiating between different kinds of the same convexity type objects, *i.e.* the bowl and the microwave in this example. Both the bowl and the microwave are containers, however due to their interactions in the specific video sequence in which the bowl is being contained in the microwave, the bowl is clustered as a container and the microwave as a containee.

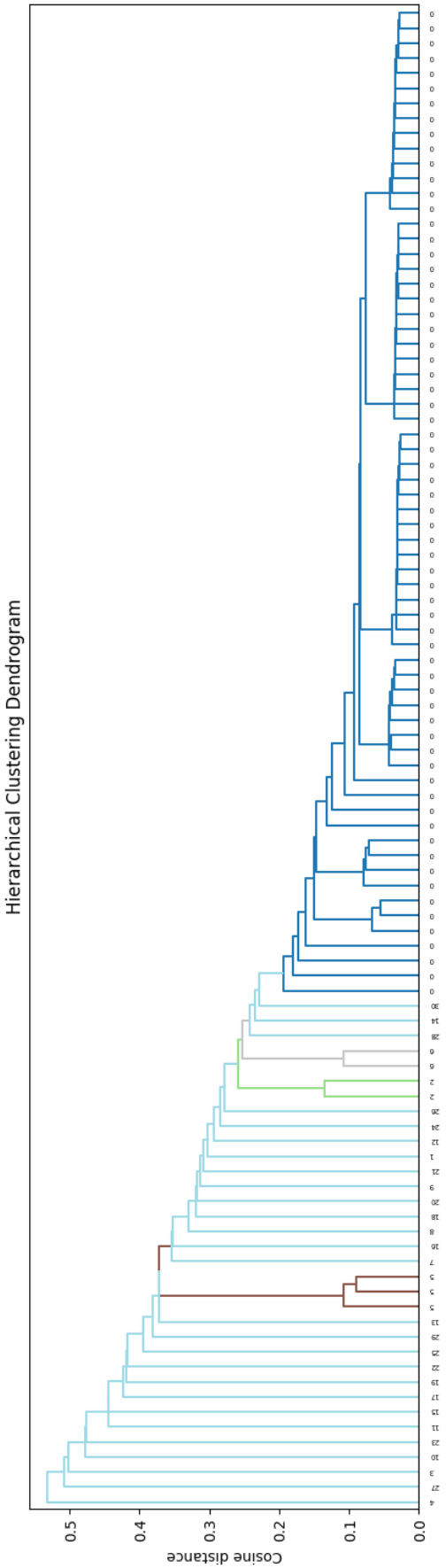
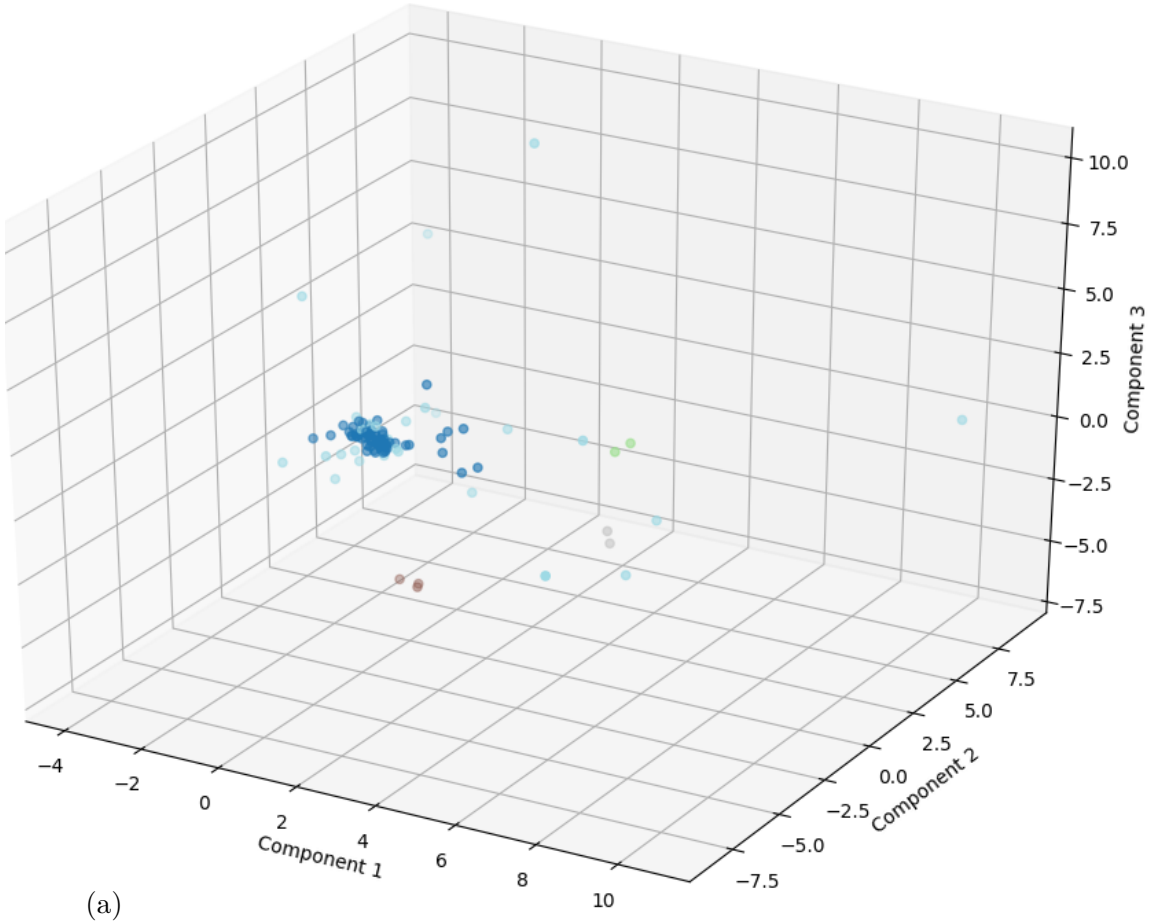


Figure 5.10: Hierarchical clustering of 100 *AGraphlets* captured from the Watch-n-Patch dataset. Dendrogram output from hierarchical clustering where the y-axis corresponds to the distance and the x-axis shows the cluster id every leaf node is assigned to; edges of the leaf nodes are colored depending on the cluster the leaf node is assigned to.



(a)

(b)

Cluster identifier	Ground truth label
0	supportable
2	holdable
5	supportable
6	holdable

(c)

V-measure	Homogeneity	Completeness
0.33	0.37	0.30

Figure 5.11: Results of hierarchical clustering of the *AGraphlet*s captured from the Watch-n-Patch video dataset from Figure 5.10. (a) Latent space with color-coded visualizations, based on the leaf color in the output hierarchy (Fig. 5.10), of the graph embeddings' location after a PCA dimensionality reduction. (b) Color-coded cluster identifiers, based on the leaf color from the output hierarchy (Fig. 5.10), mapped to the ground truth affordance labels that best describe the set of *AGraphlet*s enclosing in each group. (c) The quantitative metric scores, *i.e.* V-measure, homogeneity, and completeness, for the specific set of *AGraphlet*s.

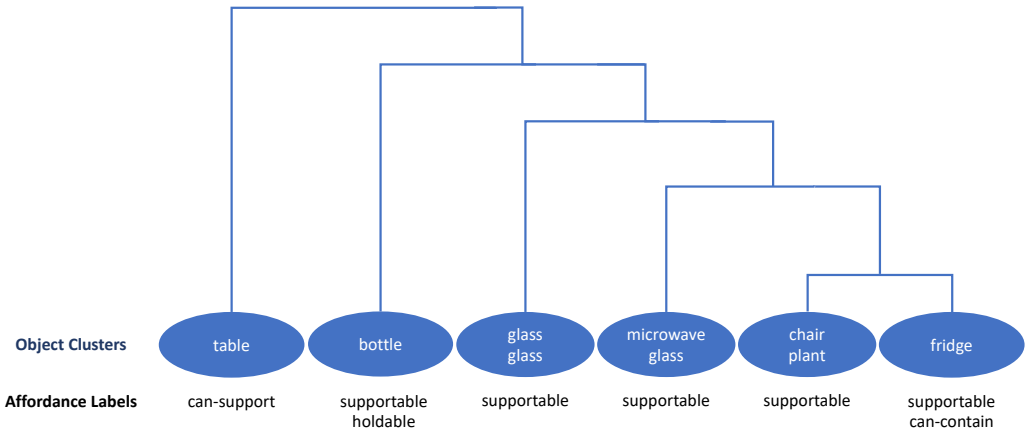


Figure 5.12: Clusters' hierarchy for a video from the Watch-n-Patch dataset.

Fine-grained categorization is also present in Figure 5.12, which illustrates the clustering outputs for a video from the Watch-n-Patch dataset. It is noteworthy that a "glass" object can appear in different clusters of the same affordance label, however each cluster captures different kinds of *AGraphlet*s, *i.e.* interactions.

Moreover, different convexity type objects can be also grouped together, if they carry the same interaction. *E.g.* in Figure 5.15, a suitcase and a trash bin object are grouped together under an affordance label 'supportable', even though a (closed) suitcase is a "convex" type object and a trash bin is a "concave" type object.

5.6 Discussion

5.6.1 Failure Cases

Object detections are retrieved from an off-the-shelf object detector, trained on the COCO dataset. However, some object configurations, present in the datasets used for the evaluation of this work, are out of the object detector's training distribution, *e.g.* the object detector's training data, for a microwave object, include only image data of a microwave with a closed door. Hence, false positive object detections are evident in cases where the primary objects change shape and visual appearance, *i.e.* opening a closed microwave. In such cases, false positive relations between detected objects are captured. *E.g.* opening the door of a closed microwave produces a detection for the main body of the microwave object as well as a different one for its door, resulting to the presence of *AGraphlet*s between the door and the rest of the objects in the scene. Such *AGraphlet*s, in which the primary object represents a part of an object

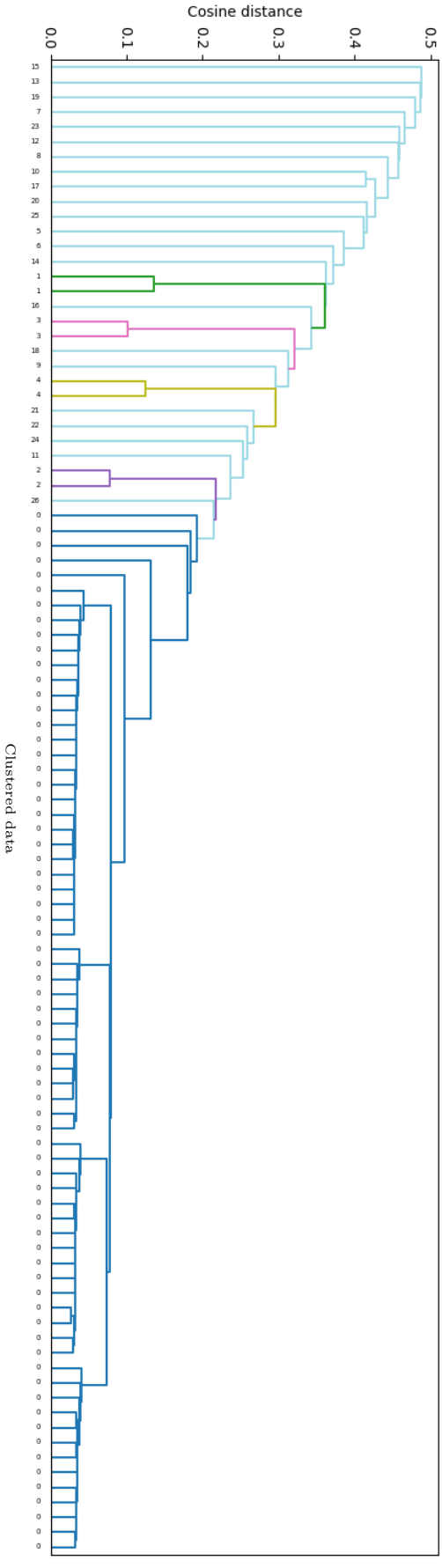
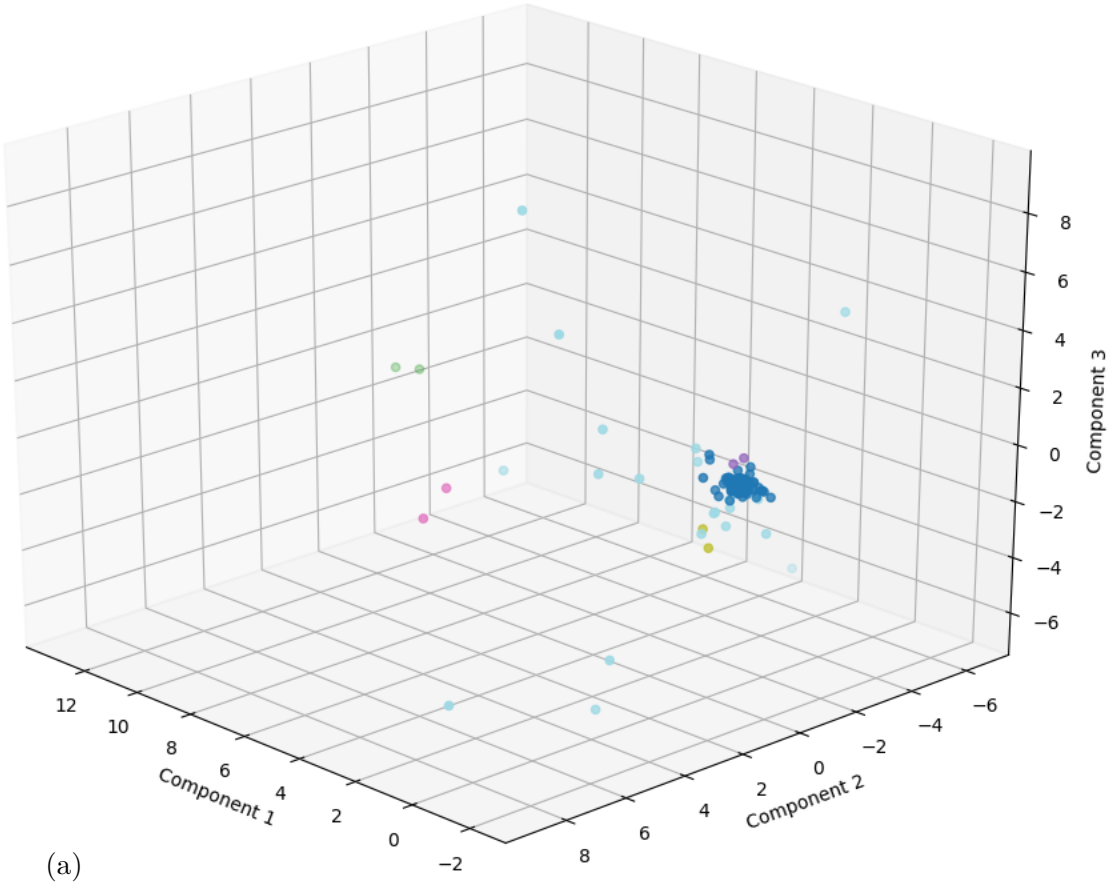


Figure 5.13: Hierarchical clustering of 100 *AGraphlets* captured from the LOAD dataset. Dendrogram output from hierarchical clustering where the y-axis corresponds to the distance and the x-axis shows the cluster id every leaf node is assigned to; edges of the leaf nodes are colored depending on the cluster the leaf node is assigned to.

(a)



(b)

Cluster identifier	Ground truth label
0	can-contain, holdable, supportable
1	unknown label
2	can-support
3	supportable
4	unkown label

(c)

V-measure	Homogeneity	Completeness
0.67	0.63	0.71

Figure 5.14: Results of hierarchical clustering of the *AGraphlet*s captured from the LOAD video dataset from Figure 5.13. (a) Latent space with color-coded visualizations, based on the leaf color in the output hierarchy (Fig. 5.13), of the graph embeddings' location after a PCA dimensionality reduction. (b) Color-coded cluster identifiers, based on the leaf color from the output hierarchy (Fig. 5.13), mapped to the ground truth affordance labels that best describe the set of *AGraphlet*s enclosing in each group. (c) The quantitative metric scores, *i.e.* V-measure, homogeneity, and completeness, for the specific set of *AGraphlet*s.

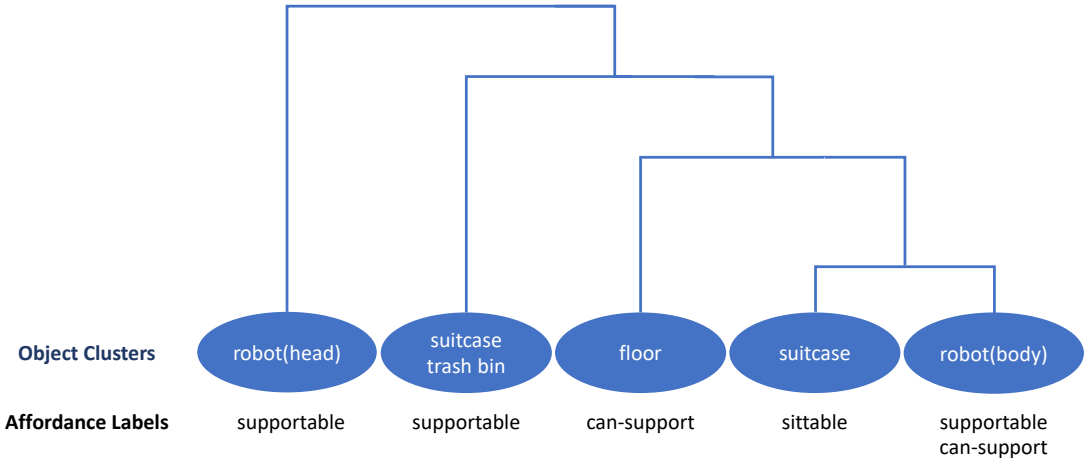


Figure 5.15: Clusters hierarchy for a video from the LOAD dataset.

in the scene, create false positive as well as false negative detections of object affordances, as the detected affordance is now correlated to the detected object part, rather than the object itself. Hence, the proposed approach is heavily reliant to the quality of object proposals.

5.6.2 Limitations

The proposed definition of spatial interactions model 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 pour" are inferred from the transition of the state Cont(bowl1, liquid) to Cont(bowl2, liquid). The proposed method is limited to only detect state-based spatial relations; hence the temporal dimension is not encapsulated in the relational representation, so affordances as "pourable", "able to pour" and "throwable" are not detectable in the current pipeline. Although 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 stirrer to be contained in a concave object.

5.7 Conclusions

In this Chapter, the problem of affordance categorization was addressed by creating embeddings of graphs of human-object and object-object interactions using the *graph2vec* network, and clustering these graph embeddings in an unsupervised way, forming groups of similar affordances in the latent space. The experiments conducted demonstrate that by exploiting depth-informed

relations produces more accurate qualitative graphs, describing the interactions between objects, resulting in more homogeneous and complete clusters of affordances, and higher V-measure scores.

The enhancement of the graph representations with additional object information, such as the size of the objects, is a future direction for expanding the possible affordances the proposed method can detect. Another avenue for future work is to enhance qualitative spatial relations, by detecting multi-object relations and enabling the construction of sequence-based relations, which allow the description of interactions that occur during a series of episodes, *e.g.* "throwing".

Chapter 6

Interaction Anticipation

6.1 Introduction

Performing long-range predictions of spatio-temporal information from video data is a challenging problem, evident in many real-world applications, such as self-driving, robot control, and human-robot collaboration, as well as perception tasks, as action prediction and object tracking.

In a human-robot collaboration scenario, predicting future interactions of objects in a scene is crucial for aiding the human, *e.g.* assisting the human when performing a physically hard task. This becomes challenging with the variability of ways the same activity can be performed by a human, as the interaction space enlarges.

In everyday-life scenarios, humans complete the same activities but each time in a different way, *e.g.* in "making tea" each one considers a different amount of time to be sufficient for brewing, *e.g.* some prefer a short brewing time lasting a couple of minutes, however others, using a jug, brew their tea for a couple of hours. When performing long-range predictions, such variations in the activities should not impact the anticipated future actions, as the final state of the world is the same. Thus, a representation of the interactions, invariant of the time duration for each holding relation, is of essence.

Predicting future interactions of a video scene is highly correlated with video frame prediction, which is the generation of future video images whilst learning scene representations from the historical ones. This approach creates an informative output, *i.e.* future image scene; however due to the nature of the data used for training, *i.e.* pixel values, the predictions do not carry any

information about the interactions taking place in the scene visualized. Moreover, predictions are constrained to the visual information used during training time. They are also limited to predicting a pre-specified number of visual frames. Predictions further in the future become blurry due to prediction uncertainty, hence long-range activities cannot be captured.

To overcome the constraints imposed by relying on the nature of video data for anticipating future interactions, high-level qualitative graphical structures are exploited, to represent object interactions present in a video, abstracting from the feature space of the image scene. This graphical representation of interactions captures the semantics of the interactions taking place and attains representation generalization across different scenes, activities, and objects. Additionally, *episode* detection is used for the creation of frame-number-independent graphs. Interactions of objects are captured using high-level qualitative spatio-temporal relations, as presented in Chapter 2.2, to achieve generalization.

Prediction from data of a single interaction can be challenging to perform as motion plays a crucial role to infer intention, *e.g.* to infer if an object is being picked up or put down a sequence of past visualizations of the scene need to be processed. For this reason, predictions take into account a sequence of past interactions, by exploiting a recurrent neural network. Spatio-temporal correlations of qualitative graphs between interactions are learned in a self-supervised way, and future qualitative graphs of future object interactions are predicted by exploiting a network architecture based on Convolutional LSTM units, due to the input's data dimensions.

The objectives of this chapter are:

- to create a 3D tensor representation that captures the information from high-level relational graphs;
- to predict future interactions in frames-number-independent time intervals, based on episode detections;
- to predict future graph representations of object interactions by employing a Convolutional LSTM-based network.

This chapter is organized as follows. Section 6.2 summarizes the related works on future interaction prediction. Section 6.3 introduces a 3D adjacency matrix representation of qualitative graphical structures which capture the information of object interactions from video data, and Section 6.4 presents the recurrent deep neural network employed for learning graph representa-

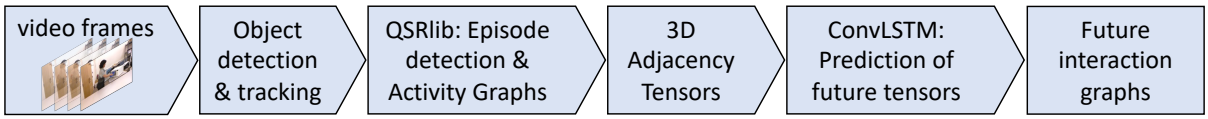


Figure 6.1: An overview pipeline of the proposed approach for future interaction graph prediction.

tions of future interactions. Furthermore, the experimental evaluation of the proposed approach is shown in Section 6.5 along with some discussion on baseline approaches from the literature in Section 6.6. Lastly, this chapter concludes with Section 6.7.

6.2 Literature Review

Some prior works focused on *short-term video prediction*, such as the ContextVP network (Byeon et al. 2018) which models contextual dependencies. Furthermore, the MCnet network (Villegas et al. 2017) acts on the motion and content of the video data, separating them into different encoder paths and performing prediction of future frames considering the observed motion. Also, inspired by "predictive coding", frame predictions with PredNet (Lotter et al. 2016) are based on the deviations of local predictions from every layer of the architecture.

Other works have focused on *long-range prediction* networks, such as the Convolutional LSTM network (Shi et al. 2015), which integrates convolutions into state transitions of a recurrent neural network. Also, action-conditioned video prediction networks (Finn et al. 2016) model pixel motion for learning physical object motion, and continual predictive network (G. Chen et al. 2022) learns from mixture world models to predict non-stationary physical environments.

Another long-range prediction network proposes a memory transition mechanism which memorizes local appearance and motion for short-term spatio-temporal predictions and exploits an attention mechanism on previous memory cells for long-range predictions (Y. Wang et al. 2018). Similar to MCnet, DRNet (Denton et al. 2017) considers two separate Encoder pathways for the object pose and the content of the video for better prediction quality. Moreover, the VPN network (Kalchbrenner et al. 2017) models the factorization of the joint likelihood of the video by estimating local dependencies of neighboring pixels. Also, two-stream recurrent neural networks are employed for capturing the different frequency domain information (J. Xu et al. 2018), as well as to perform reconstructions of the current frames and predictions of future ones (Srivastava et al. 2015).

Thus, works in the literature relate the problem of interaction anticipation with future video frame prediction, by solely relying on the visual data of the video image frames. Their pixel-level predictions of future interactions have high uncertainty after a few frames causing blurriness of the output. Moreover, learning from visual features constraints the model's generalizability across different domains as the features learned are based on the visual appearances present in the dataset used for training the models. This chapter address these limitations by presenting a novel approach for addressing the problem of spatio-temporal anticipation considering object interactions from real-world video data.

6.3 Interaction Sequence Modeling

As graphical structures are able to capture high level information and achieve generalization across different domains, graphs of qualitative relations are exploited to represent activities between objects in a video scene considering their spatio-temporal interactions.

For this purpose, a variant of *Activity Graphs* (Sridhar et al. 2010a; Sridhar et al. 2010b) (*vAG*s) is used to represent entity, *i.e.* object, interactions present in a video. These graph representations (g) comprise two layers of vertices where each layer consists of a single type of node and only nodes in adjacent layers can be connected with each other. The bottom (object) layer contains the set of vertices representing the interacting entities of the video, and the top (spatial) layer consists of the vertices with the spatial relations describing the spatial interaction of the entities. The spatial relations captured are:

- for every pair of objects, the relationships from the *Region Connection Calculus* (RCC5) (D. A. Randell et al. 1992; Anthony G Cohn et al. 1997) which consist of the relations: "discrete" (DR), "partially overlapping" (PO), "proper part" (PP, PPi), and "equal" (EQ),
- for every pair of objects, the relationships from the *Qualitative Trajectory Calculus* (QTC) (Van de Weghe, Anthony G Cohn, De Tre, et al. 2006; Delafontaine et al. 2011) which contains the relations: "-,-", "-,0", "-,+", "0,-", "0,0", "0,+", "+,-", "+,0", and "+,+"; where the pair "α, β" represents the relative motion of each object towards the other and "+" means motion away, "-" means motion towards, and "0" means no relative motion,
- for every object, a binary state of *Moving or Stationary* (MoS) ,
- and the *Cardinal Direction* of motion (CarDir) (Frank 1991) for every moving object in

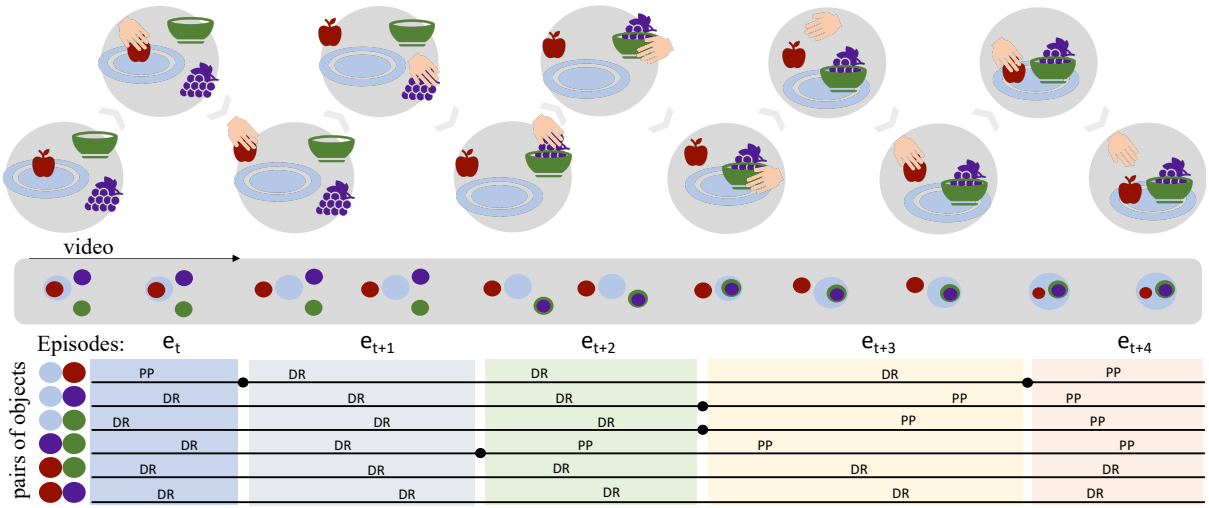


Figure 6.2: Episode detection in a demo video with several color-coded objects. For simplicity only the RCC5 relations are visualized.

the scene, that corresponds to the set of relations: "north" (N), "north east" (NE), "east" (E), "south east" (SE), "south" (S), "south west" (SW), "west" (W), "north west" (NW), and "equal" (EQ).

The maximum period of time throughout which a spatial relation between the video entities occurs, whilst before and after that time a different spatial relation holds, is an *episode* (e), and multiple episodes define the sequence of spatial relations obtained in every interaction. Figure 6.2 presents the episodes detected in an example video.

6.3.1 Tensor Representation

At training time, for every video, object proposals are utilized to define entities and a *vAG* (g) is extracted, representing each detected episode and consisting of all the objects involved. *E.g.* in Figure 6.3 episode e_{t+4} is represented by *vAG* g_{t+4}. The temporal information for every *vAG* is expressed by ordering them in temporal episode-detection order.

A 3D tensor representation is exploited for the *vAG*s (Fig. 6.3). Each tensor describes the spatial relationships holding in an episode. Hence, a sequence of tensors carries all the interactions present in a video. A *vAG* tensor $T \in \{0,1\}^{O \times O \times R}$ is based on the construction of a 3D adjacency matrix between all entities and spatial relations of an episode, where O is the number of entities and R the number of relations. Thus, the values in T are assigned based on Equation 6.1 where o_1, o_2, and r are the locations of the two objects and the location of the

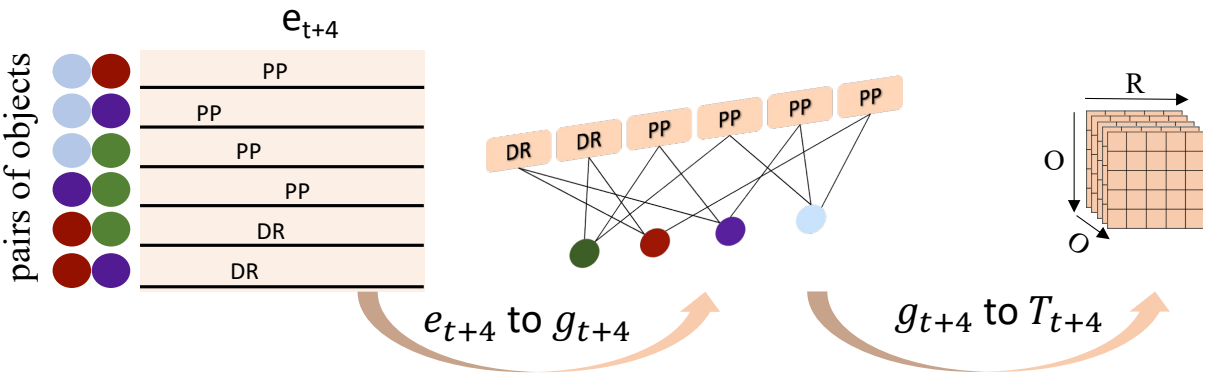


Figure 6.3: Tensor representations from episode-based qualitative graphs. Every tensor captures the spatio-temporal information of a single episode.

relation ($relation_r$) respectively, in T.

$$T[o_1, o_2, r] = \begin{cases} 1 & \text{if } relation_r(obj_{o_1}, obj_{o_2}) = \text{True} \\ 0 & \text{if } relation_r(obj_{o_1}, obj_{o_2}) = \text{False} \end{cases} \tag{6.1}$$

Since the size of the tensor is static, it does not change depending on the number of detectable objects; some object rows will be filled with zeros if fewer than O objects are detected. A zero value in a detected object specifies that the specific relation between that object and another is not present, whereas a zero value for a non-detected object means that the object is not present. To explicitly differentiate these two cases, for every relational set an extra relation "not applicable" (N/A) is added, which applies to all non-detected objects. Also, O is selected to be sufficiently big so the number of detected objects does not exceed the tensor's size.

6.4 Qualitative Interactions Prediction Network

For capturing long-range dependencies in multi-dimensional tensors for representing qualitative spatio-temporal information Convolutional LSTM units (ConvLSTM) (Shi et al. 2015) are exploited. Convolutional LSTMs were introduced as an extension of the Fully Connected LSTM network (FC-LSTM) considering convolution structures in the input-to-state and the state-to-state transitions. All features are represented by 3D tensors with dimensions (height $\times$ width $\times$ channels), insstead of vectors, and the matrix multiplications are replaced with tensor convolutions, compared to the LSTM units. The parameters of a ConvLSTM are the input weights $W_x \in \mathbb{R}^{K \times K \times P}$ and the recurrent weights $W_h, W_o \in \mathbb{R}^{K \times K \times F}$ with K, P, and F denoting the kernel size, the number of channels of the input $X_t \in \mathbb{R}^{H \times W \times P}$ and the hidden

states $H_t \in \mathbb{R}^{H \times W \times F}$ respectively. The key equations of ConvLSTM are:

$$
\begin{aligned}
i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i) \\
f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f) \\
C_t &= f_t \circ C_{t-1} + i_t \circ tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \\
o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o) \\
H_t &= o_t \circ tanh(C_t)
\end{aligned}
\tag{6.2}
$$

where $\sigma(\cdot)$ represents a sigmoid function, "$*$" denotes a convolution, "$\circ$" the Hadamard product, and $i_t, f_t, o_t, C_t \in \mathbb{R}^{K \times K \times F}$ are the input gate, forget gate, output gate and memory cell respectively. The input to the network is of dimension $(M \times H \times W \times P)$, where M is the number of samples, $H = W = O$ represent the number of detectable objects, and P is the set of the spatial relations captured (29). Also, the output tensor is of dimension $(H \times W \times F)$, where $H = W = O$ and F is the number of relations to be predicted. The output is selected to describe spatial positions of the objects based only on RCC5 since such information is considered as the most useful for many real-world applications[1]. Thus, $F = 6$ is set, where 6 is the number of RCC5 relations including an additional N/A relationship for non-detected objects.

At deployment time, this network processes incrementally a sequence of *vAG*s for predicting the future state of objects' interactions. The networks updates its internal recurrent state for every episode of interactions, accumulating in that sense past information and enhancing its future predictions. The complete representation of the input/output tensors and the proposed pipeline are shown in Figure 6.4. The input data capture the information of object interactions for all the relations (RCC5, QTC, MoS, CarDir), though the output represents the spatial interactions from the RCC5 set of relationships only.

6.4.1 Self-supervised Training

The network is trained in a self-supervised way, by exploiting the sequential nature of the tensor data. More specifically, at timestep t the input comprises of the tensor data at time t (T_t) and $t-i$ (T_{t-i}), where $i \in \{1, \ldots, b\}$, and where b represents the batch size. The prediction at time t (T_{t+1}) is compared against the ground truth tensor at time $t+1$, to update the model's weights.

[1]The model would not reach convergence when QTC, CarDir and MoS relations were considered in the output space.

6.4.2 Training Loss

Due to the nature of the data, *i.e.* T are binary sparse tensors, the task is to correctly predict the correct tensor as a multi-class classification problem for every pair of objects between the different spatial relations captured. Moreover, since the appearance of relations in the data is imbalanced, to avoid overfitting on the most dominant relations, a weight is assigned to every relation class. Hence, for updating the weights of the network in every iteration a weighted categorical cross-entropy loss function is minimized. The loss function is defined as:

$$\mathscr{L} = \frac{1}{M} \sum_{k=1}^{K} \sum_{m=k}^{M} w_k \cdot y_m^k log(h_\theta(x_m, k)) \tag{6.3}$$

where M represents the number of training examples, K is the number of classes, w_k is the weight of class k, h_θ represents the model with neural network weights θ, y_m^k is the target label for the training example m of class k, and x_m denotes the input of the training example m. The weight of each class k (w_k), where $k \in K$, is set percentage-wise depending on the overall detection of each one across the whole dataset ($w_k = \frac{n}{n_k \cdot |K|}$, where n_k represents the number of samples in class k). Hence a relation that appears often will have a low weight, whereas a relation that appears rarely in the dataset's interactions will have a higher weight.

6.4.3 Model Architecture

Inspired by the model architecture proposed by Shi et al. 2015, the proposed network comprises of a series of layers of ConvLSTM modules with 128, 64, 64, and 32 hidden states outputting to a 6 channeled tensor (Fig. 6.4). Hence, the output tensors are of dimension ($O \times O \times 6$) and the input tensors are of dimension ($O \times O \times 29$), where O represents the number of detected objects. Furthermore, the kernel size is set to $(1, 1)$ as every value in the tensor is independent of its neighbors, and the weights are initialized based on the LeCun uniform distribution (LeCun et al. 2012).

6.4.4 Training Hyper-parameters

During training, the Adam optimizer was employed for the update of the weights. The learning rate started from 0.01 along with a scheduler to reduce the value of the learning rate every 1000 epochs by a factor of 0.1. The model was trained until convergence for a maximum of 2.5k

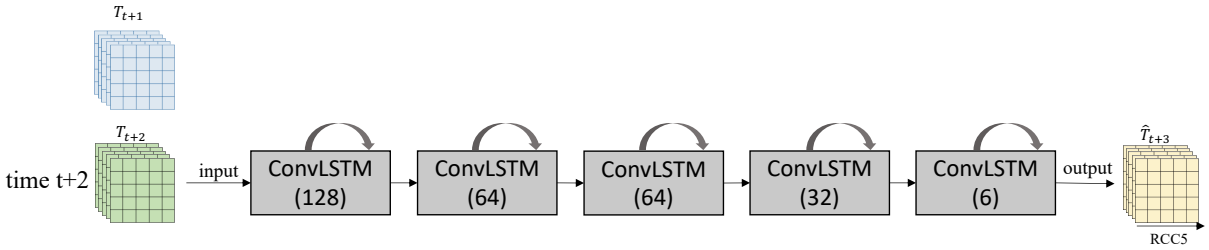


Figure 6.4: Interaction prediction network based on ConvLSTM units for predicting the future RCC5 relationships from qualitative spatio-temporal tensor representations.

epochs. Moreover, the batch size was set to 5 considering the minimum number of episodes captured from a single video[2]. Thus, no batch disturbs the temporal ordering of the data by shuffling or concatenating data from different videos. Also, a Lasso regularization term was added, of $\lambda|w|$ with $\lambda = 1e-4$ in the loss function to avoid over-fitting of the model.

6.5 Experimental Evaluation

6.5.1 Dataset

The proposed approach was trained and evaluated on the CAD-120 dataset (Koppula et al. 2013) exploiting the ground truth bounding boxes of object positions. The CAD-120 dataset comprises 120 RGB-D sequences of frames of everyday-life activities, capturing human-object interactions in various scenes, *e.g.* office, kitchen, *etc.* From every video of the dataset, tensor representations of the object interactions were created by considering the input relations extracted using the QSRlib library (Gatsoulis et al. 2016). Due to the static size of the tensors the maximum number of detected objects (O) was set to be 10, which is adequate for capturing all the object interactions in the employed dataset. Also, in every epoch, the tensor data are shuffled along both the object axes so no correlation between the rows is learned.

6.5.2 Evaluation

This experimental evaluation of the proposed approach aims to encapsulate how well the proposed model can predict qualitative spatial relations, which describe the object interactions taking place in the next episode. The predicted outputs comprise of probability distributions across all the output relations for each pair of objects. However, since open-ended scenarios as hard to specify and evaluate, only the top-1 predictions are evaluated.

[2]Video data with fewer than 5 episodes are not considered. Also, remaining episodes after batching, are not considered if they are fewer than 5.

Model	J.I. (↑)	W.C.E. (↓)	C.Acc. (↑)	F1 (↑)	Training parameters num. parameters	batch
S_1: RCC5	0.4228	0.9765	0.9198	0.4504	164,904	4
S_2: RCC5+QTC	0.5747	0.7453	0.9454	0.5921	170,024	5
S_3: RCC5+QTC+CarDir	0.6329	0.8512	0.9469	0.6472	175,144	5
RCC5+QTC+CarDir+MoS	**0.6477**	**0.6133**	**0.9621**	**0.6659**	176,680	5

Table 6.1: Quantitative results of the ablation study experiments on the test set.

Experiments were conducted using different qualitative spatial calculi to evaluate how the incorporation of each relational set helps improve the predicted output. The predicted tensors represent the future interactions of the next episode, by considering the current and a set of previous episodes. For the studies S_3 and S_2 the batch size was set to 5 whereas for S_1 to 4, due to the smaller number of episodes detected, fewer spatial relations denote fewer episodes.

Inspired by the evaluation metrics in the *instance segmentation* literature, for quantifying the overlap of 1s in the predicted tensors over 1s in the ground truth tensors of RCC5 spatial interactions, one of the metrics employed is the *Jaccard similarity index* (J.I.) (Eq. 6.4) for multiple classes, which considers the number of classes (K) and the true positives (TP_k), false positives (FP_k) and false negative (FN_k) for every class (k).

$$\mathscr{J} = \frac{1}{|K|} \sum_{k=1}^{K} \frac{\mathrm{TP}_k}{\mathrm{FP}_k + \mathrm{FN}_k + \mathrm{TP}_k} \tag{6.4}$$

Moreover, additional reported metrics are the *categorical accuracy* measure (C.Acc.), the *F1-score* (Van Rijsbergen 1979) (F1), as well as the *weighted cross-entropy* loss value (W.C.E.) for every experiment. The proposed method was evaluated on 25% of randomly-picked unseen video data and the results in Table 6.1 demonstrate that the proposed approach achieves the best results in all reported metrics[3]. More specifically, the proposed approach combining the information of 1) the spatial location, 2) the relative motion, 3) the absolute motion, and 4) the direction of the absolute motion of the objects, can achieve an increase of the Jaccard index score of 2%, 13% and 53% compared to the the studies S_3, S_2 and S_1 respectively. Moreover,the results demonstrate an increase of the categorical accuracy of 1.6%, 1.8% and 4.6%, as well as an increase of the F1-score of 2.9%, 12.5% and 47.8% compared to the studies S_3, S_2 and S_1, respectively. The weighted cross-entropy loss value shows a significant improvement of 27.9%,

[3]↑ indicates "highest is best" and ↓ indicates "lowest is best"

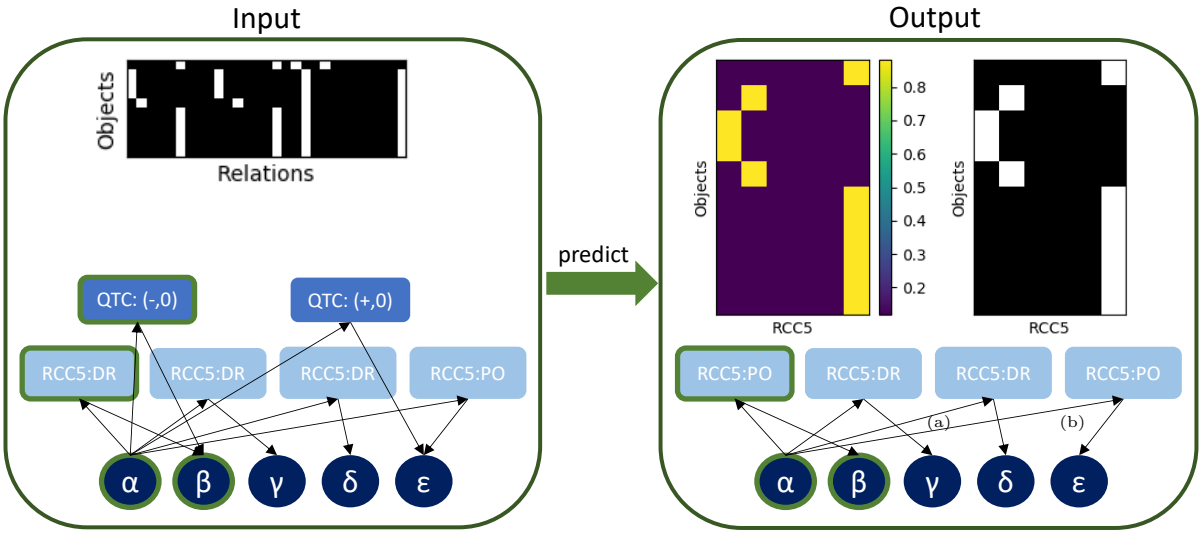


Figure 6.5: Qualitative results in an example case for the interactions with object α. Output matrix (a) corresponds to the network prediction, whereas matrix (b) represents the ground truth relations. White cells contain the value 1 and black cells the value 0. Yellow and purple cells are the predicted values closer to 1 and 0, respectively.

17.7% and 37.2% compared to the studies S_3, S_2 and S_1, respectively, since it considers the imbalance of the data by applying a weight at each relation. Furthermore, these results were attained with a growth of 7.1%, 3.9%, and 0.9% in the model size over the studies S_1, S_2 and S_3, respectively.

Some qualitative results are illustrated in Figure 6.5 along with the corresponding *vAG*s, with the pair-wise relations, for one of the interactive objects (object α). Figure 6.5 shows a visual representation of a two dimensional snap shot of the model's input, along with the model's prediction and corresponding ground truth tensor, of the relationships between all objects and object α. For simplicity only the graph information for RCC5 and QTC relations is visualized, omitting the MoS and CarDir relations from the input tensor.

It is evident that the motion information (QTC:(-,0)) as well as the direction of motion, signify that object β is moving towards object α. Hence, in the predicted *vAG* a PO RCC5 relation holds between objects α and β. The values of the predicted tensors are binarized by setting the switch point to 0.5. Thus, values greater than 0.5 are considered as 1.0 and 0.0 otherwise. Hence, by binarizing the predicted tensor the prediction for this example maps exactly to the ground truth.

6.6 Discussion

6.6.1 Baseline Comparison

This work presents a novel research direction towards predicting future interactions between objects, in short- as well as long-range activities by exploiting episodes rather than a frame-by-frame representation. This section introduces some qualitative results of related and state-of-the-art works on predicting future visual instances of scenes. The visualized outputs showcase the impact of blurriness on the predicted frames, as well as the importance of having number-of-frame independent predictions.

Qualitative Results of FP-LSTM

Similar to the proposed approach, FP-LSTM (Srivastava et al. 2015) employs a recurrent network architecture based on LSTM modules. Predictions comprise 10 image frames based on an input of 10 sequential input frames[4]. Figure 6.6, Figure 6.7, and Figure 6.9 illustrate some qualitative results of the predictions from the trained model in Moving MNIST, UCF101 (Soomro et al. 2012), and CAD-120 datasets, respectively. The model was trained using one layer and two layers of LSTMs, *i.e.* stacked ontop of each other. More qualitative results can be found in Appendix C.

From these figures it is apparent that two layers of LSTMs produce less blurry results in the last predicted frames. However, visual instances are still noisy and interactions are not clearly distinguishable. Specifically, in the Moving MNIST dataset the digits are not recognizable[5] and in the UCF101 dataset the image frames are too noisy to distinguish the interaction taking place. The results in CAD-120 look more promising in terms of noise. Nonetheless, a prediction length of only 10 frames is not long enough to allow to capture the future evolution of spatio-temporal interactions as few such changes occur within a 10 frame interval.

Qualitative Results of CPL

State-of-the-art work CPL (G. Chen et al. 2022), performs future predictions by employing mixture world models. Figure 6.10 illustrates the predicted outputs of CPL on the "boxing"

[4]Training was performed for a 5 and 15 frames predictions on the Moving MNIST dataset, with the input sequence changing accordingly. The qualitative outputs of their predictions were worse than the predictions of 10-frame sequences, as the digits were un-recognizable from the first predicted frames.

[5]Since Moving MNIST depicts only moving digits, a possible enhancement, to acquire less blurry predicted images, would be to extrapolate the input data to include also the video data in a backwards sequence.

Figure 6.6: FP-LSTM predictions on the Moving MNIST dataset for moving digits 5 and 6. The model was trained with a single FP-LSTM layer (l_1) and a double layer (l_2) separately. Ground truth image frames are shown in the "gt" rows.

Figure 6.7: FP-LSTM predictions on the UCF101 dataset for a scene without a lot of spatio-temporal variations between the visualized frames of the video sequence. The model was trained with a single FP-LSTM layer (l_1) and a double layer (l_2) separately. Ground truth image frames are shown in the "gt" rows.

Figure 6.8: FP-LSTM predictions on the UCF101 dataset with some spatio-temporal variation between the frames in the presented video sequence. The model was trained with a single FP-LSTM layer (l_1) and a double layer (l_2) separately. Ground truth image frames are shown in the "gt" rows.

Figure 6.9: FP-LSTM predictions on the "making cereal" activity of the CAD-120 dataset. The model was trained with a single FP-LSTM layer (l_1) and a double layer (l_2) separately. Ground truth image frames are shown in the "gt" rows.

activity from KTH Action dataset (Schuldt et al. 2004). Moreover, Figure 6.11, Figure C.10, and Figure C.11 present the qualitative outputs of CPL on the "making cereal", "taking medicine", and "taking food" activities on the CAD-120 dataset, respectively.

Two frames are used as an input to the CPL model and predictions go up to 20 frames. The produced images from the CAD-120 dataset are black and white colored since the CPL model is trained on black and white image data, *i.e.* it is using only 1 channel at the input for the image data.

The results illustrated in these figures indicate that after the 11th frame the output images become noisy, thus interactions are not detectable. Specifically, in Figure 6.11 it is evident that even from the first predicted frames it is not possible to predict future interactions and understand what is happening in the scene, due to pixel-level uncertainty. In the final frames, the human in the scene is not visible any more. Furthermore, the CAD-120 dataset requires longer sequences of frames to be predicted to visualize the spatio-temporal differences during an activity.

Differently from the aforementioned baselines, the proposed method does not have a maximum number of frames that predictions can be performed, since predictions are based on episodes. Moreover, uncertainty in the output is presented as multiple relations having high values in the object-pair axis of the predicted tensor. Higher tensor values indicate a more probable relation. Thus, even in noisy scenarios, the proposed approach can provide a meaningful predictions.

6.6.2 Limitations

The proposed approach considers high-level information of interactions between objects, aiding generalization across different domains. However, object affordances, shape and status, *e.g.* open/closed microwave, are not captured though carrying important information about possible future interactions. Also, the input data in the proposed method are sparse binary tensors. Such kind of data make the training process hard to converge and the input data challenging to fuse with other kind of data[6], *i.e.* data from different modalities. In addition, the proposed method considers a fixed size of objects to be detected in a scene. Thus in a cluttered scene, where several interacting objects are not detected, the appropriate future interactions are not

[6]Training convergence of the model was not possible in experiments conducted with an implementation of an early-fusion, *i.e.* concatenation, of the binary tensor data and the visual embeddings of the detected objects, extracted from the Mask R-CNN framework.

Figure 6.10: CPL predictions for "boxing" video activity of the KTH Action dataset.

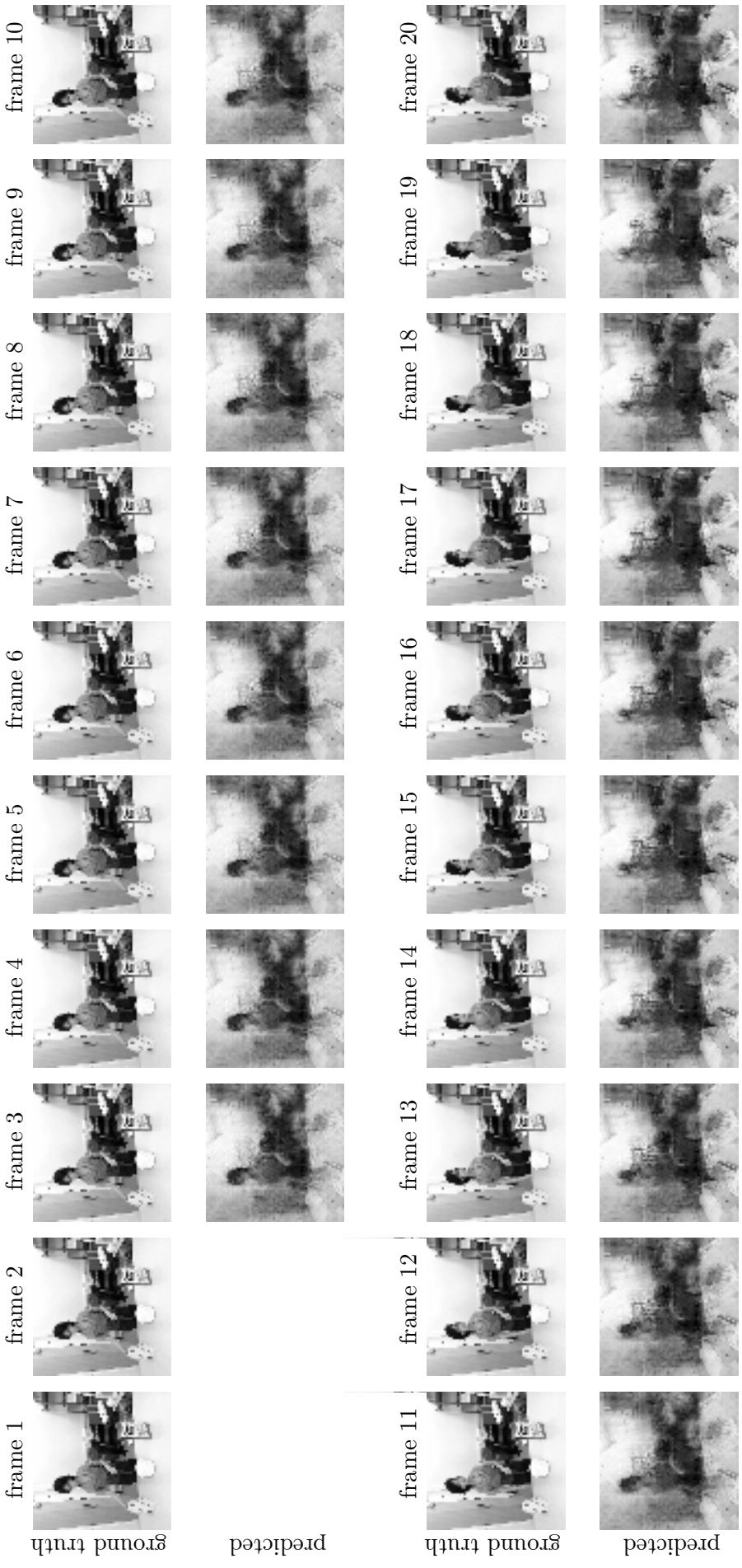


Figure 6.11: CPL predictions on "making cereal" video activity of the CAD-120 dataset.

predicted.

Moreover, dependencies of interactions between multiple objects in an episode are not captured. *E.g.* in a scenario where a microwave object is supported by a table object, the prediction of the relation of a bowl object being contained in the microwave should also infer the relation that the bowl is supported by the table.

Furthermore, though uncertainty in the output data is handled so that the predictions are meaningful interactions, the aforementioned dependency of interactions between multiple objects should also be represented on the predicted tensor data.

6.7 Conclusions

This chapter presented a novel approach for solving the task of interaction anticipation whilst exploiting high-level qualitative spatial representations and training a ConvLSTM-based network in a self-supervised way. The qualitative graphical representations used for capturing object interactions push the limits of interaction anticipation towards generalizability across various domains and predictions which are independent of the time-duration. The evaluation results demonstrate that exploiting a rich set of high-level relations is a promising direction for predicting future spatial interactions, whilst not being frame-number dependent.

Chapter 7

Conclusion

This thesis has presented a new method for learning affordance categories of previously unseen objects in an unsupervised way; it has also introduced a novel approach for predicting future object interactions in a video scene. A major challenge is attaining generalization across different objects, scenes, activities, and human agents. Qualitative spatio-temporal relations were exploited to abstract from the primitive continuous space of interactions to create generalizable representations of object interactions, in different domains.

This chapter is organized as follows: the contributions of this research are presented in Section 7.1, followed by a discussion about the limitations in Section 7.2 and some future work in Section 7.3. The thesis concludes with some final remarks in Section 7.4.

7.1 Contributions

This research makes the following contributions:

1. a novel depth-informed mereotopological set of relations, that captures complex spatial relationships, *i.e.* "supporting", "containing", and enables the distinction between occlusions and interactions;
2. a newly introduced RGB-D video dataset that captures everyday-life activities with human-object interactions, expanding the domain of object affordances;
3. an unsupervised learning method on qualitative graph embeddings for creating categories of object affordances from previously unseen objects;

4. a self-supervised method based on a Convolutional LSTM model for predicting number-of-frame-independent future qualitative graphs of spatial object interactions.

7.2 Limitations

This section presents the limitations of this thesis, which focus mainly on the modality for extracting object proposals, and the representation of interactions between objects.

Firstly, spatial interactions between objects are subject to the detection of objects on the input visual data. A pre-trained deep neural network is exploited to provide proposals of object detections. To allow generalizability across different object sets, a computer vision technique for detecting low-level features needs to be employed.

Moreover, the relations captured in this work do not encapsulate the temporal dimension. Hence, complex interactions that comprise of a sequence of state-based spatial relations are not captured.

Furthermore, interactions with dependencies between multiple objects are not considered in this thesis. *I.e.* interactions that are derived from interactions between a set of objects, *e.g.* for a stack of books on a table, the table supports each one of these books.

7.3 Future Work

This section presents possible improvements of the different aspects of this research.

7.3.1 DiSR Definition and Detection

The DiSR definitions proposed in Chapter 3 are an approximation of the English meaning they are referring to. A future research direction is to refine these relationships to more closely correspond to the semantics of the English words. *E.g.* "support" in English is used to describe the relation of an object holding an object at a specific height *e.g.* the table supports the bowl, the nail supports the frame, or an object holding another object from falling over *e.g.* the book-holder supports the books.

Moreover, there can be situations where the shape of a "concave" type object, allows other objects to interact with a non-concave area, with similar depth information as in the cavity region. The proposed approach is limited to objects which do not have non-concave regions

with similar depth as in the cavity area. Hence, an enhancement of the DiSR definition of Cont is possible, to address this limitation by considering the location of the concave region. In addition, a more generalizable approach to employ the relative shape/size of the object as well as its distance from the camera can aid for considering a dynamic convex depth threshold. Thus, the detection of Cont will not rely on a predefined depth threshold value.

Also, the On relation, is defined in the 2D space, not allowing the generalization across a wide variety of actual "ontop" situations to be detected, *e.g.* wearing a hat *on* the head wear the sides of the hat exceed the head diameter. Thus, an event-based definition usage can aid in defining relations based on event occurrence (Siskind 2001; Siskind 1994).

Furthermore, the detection of the DiSR relations relies greatly on the camera view point, since for a Cont relation to hold is necessary to detect the concave curve on the object of interest. It can be the case that a containment holds but it is not visible from the specific camera angle. Thus, camera views from different camera angles can advance the DiSR detection algorithm.

7.3.2 Definition of Relations for Affordance Detection

The proposed method for categorizing previously unseen objects depending on their affordances in Chapter 5 employs definitions of spatial interactions that model spatial states of the interactive objects, *e.g.* a cup is on the table. These defined spatial relations are limited to only detect state-based interactions, *e.g.* "grab", "lean", "contain", "support", "on top", *etc.*

However, in the real-world there exist affordances of objects that are derived from transitional interactions, *i.e.* modeling a sequence of states, *e.g.* the "pourable" and the "able to pour" affordances are inferred from the transition of the state Cont(bowl1, liquid) to Cont(bowl2, liquid).

Moreover, a sub-category of these state-sequence interactions are the ones in which one of the entities involved can alter to a different one during the transitional interactions, *e.g.* the "throwable" affordance is inferred by the C(ball, hand) to DC(ball, hand) to C(ball,_), where in the last detection of the C relation one of the entities involved can be the same hand as in the first two relations, can be a hand from another human agent, or can even be a completely different entity, *e.g.* the floor.

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 the "stirrable" affordance requires both a liquid and a stirrer to be partially contained in a concave object.

Furthermore, in the current pipeline, no information about the involved objects was exploited. Nevertheless, information such as the relative size of the objects can enhance affordance detection, *e.g.* a ball, which is much larger than a bowl, cannot be contained in that bowl. To supplement the graph representations with such additional information, one could incorporate object embeddings instead of object identifiers in the V_{ent} layer of the AGs.

7.3.3 Enhanced Interaction Representation for Interaction Anticipation

Chapter 6 proposed a novel method for predicting future object interactions in everyday-life activities, considering solely a set of high-level qualitative spatial relations. These relations captured topological and motion relational changes between objects in the scene.

However, objects' interactions do not solely rely on the previous interactions that have taken place but on the visual features of the objects as well, *e.g.* shape, state *e.g.* closed microwave, opened microwave, and relative size of the objects. Extracting feature vectors from a hidden layer from the object detector used, or creating a vector/tensor representation of low-level features computed by using a Computer Vision algorithm, *e.g.* HOG, can form the visual features of the objects. To enhance the interaction predictions with this kind of information, visual embeddings of the objects need to be incorporated in the input of the Convolutional LSTM model used for training and inference. A promising future direction for incorporating such information in the proposed model is the integration of some initial layers for learning object representations or creating embeddings of the input data before performing sequence learning. This will enable the input data to be easily fusable with data from other modalities, allowing at the same time convergence. In addition, exploiting graph convolutional layers to create learnable object representations, might facilitate the prediction of future interactions for a dynamic number of detected objects in a scene.

Morever, the incorporation of DiSR relations should also be considered in future work of this project, along with additional qualitative spatial relations. Thus, enhancing the input domain space with come complex qualitative spatial relations.

Furthermore, the proposed approach can aid human-robot interaction scenarios since prediction

of future interactions can facilitate human-robot collaboration, *e.g.* in a kitchen scenario where a human agent needs to put a big tray in the oven, predicting the interaction/action of a robotic agent to open the oven door. Thus conducting real-world experiments, using a robot agent is crucial to investigate the impact of the proposed approach in a real-world environment. In these experiments, a robotic agent will have the goal of completing an activity initiated by a human.

Another interesting research direction is the creation of visual representations of the predicted outputs, *i.e.* creating an image representing the prediction of the qualitative graph relations of the future interactions. For this task a greater enhancement of the exploited models should be considered. The object and scene semantics need to be learned. The image describing the latest state of the scene needs to be fused with the predicted graph of future interactions to output the visual representation of the future scene interactions. Thus, the achievement of visual representations, *i.e.* image data, in various future time intervals could greatly advance the research direction of future video frame prediction.

7.4 Concluding Remarks

In conclusion, this thesis has presented a novel approach for learning to categorise object affordances in an unsupervised way by solely relying on qualitative information describing the objects' interactions. To provide better interaction descriptions and representations, a novel set of qualitative spatial relations has been introduced. This relational set is able to detect more complex spatial relations than existing ones, by incorporating information of the objects' shape to infer relations. Moreover, this thesis introduces a new research direction for predicting future interactions by exploiting qualitative spatial graph representations of interactions.

A video dataset capturing human-object interactions in everyday-life activities has been made publicly available to the community, and the newly presented set of relations has been provided as an open source library, which will be incorporated in the QSRlib library. Finally, the work presented in this thesis aims to inspire new research directions in learning interactions and predicting interactions through graph data extracted from videos, by utilizing more complex information, *e.g.* relations, rather than pixel values. It also aims to motivate research on the semantics of interactions and understanding the purpose of human actions.

Appendix A

Clustering Validation Indices

For completeness, the internal clustering validation indices are presented. The notation employed in the rest of this chapter is based on the notation exploited in Section 2.4.1.

A.1 Internal Validation Indices

From the literature, validation indices are also categorized based on the way they combine the cluster *cohesion* and *separation* to compute the quality measure. Thus, cluster validity indices can either be *ratio-type* (Dunn 1973; Davies and Bouldin 1979; Caliński and Harabasz 1974; Bezdek and Pal 1998; Halkidi and Vazirgiannis 2001; Chou et al. 2004; Gurrutxaga et al. 2010; K. R. Žalik and B. Žalik 2011), *summation-type* (Rousseeuw 1987; Saitta et al. 2007), or *based on graphical representations* (Pal and Biswas 1997).

Experimental studies have shown that ratio-type cluster validity indices perform better than do other types of indices (Kim and Ramakrishna 2005). Thus, this section focuses on presenting the ratio-type and some widely used summation-type indices. Table A.1 summarizes the ratio-type metrics for internal validation, which are detailed below.

Dunn Index

Dunn index (Dunn 1973) represents a quantitative index of separation among the disjoint finite subsets of a partition of a set X. This well-established ratio-type index identifies the subsets of the partition which are well separated and compact. It considers the cohesion of the points in the subset according to the smallest distance between two points in it, as well as the separation

Table A.1: Ratio-type metrics for internal cluster validation.

Metrics	Cohesion estimation	Separation estimation
Dunn index (Dunn 1973)	Estimated by the nearest neighbour distance.	Estimated by the maximum cluster distance.
Callinski-Harabasz index (Caliński and Harabasz 1974)	Estimation based on the distance from the points in a cluster to its centroid.	Based on the distance form the centroids to the global centroids.
Davies-Bouldin index (Davies and Bouldin 1979)	The distance of the points in the cluster from its centroid.	Based on the Minkowski metric of the centroids.
Generative Dunn indices (Bezdek and Pal 1998)	Variations of cohesion estimator.	Variations of the separation estimator.
CS index (Chou et al. 2004)	Estimated by the cluster diameters.	Estimated by the nearest neighbour distance.
COP index (Gurrutxaga et al. 2010)	Based on the distance from the points in a cluster to its centroid.	Based on the furthest neighbour distance.
SV-index (K. R. Žalik and B. Žalik 2011)	Evaluated by the nearest neighbour distance.	Estimated from the distance of the border points of a cluster to its centroid.
S_Dbw index (Halkidi and Vazirgiannis 2001)	Based on the standard deviation of the set of objects in the dataset, as well as the standard deviation of the partition.	
OS-index (K. R. Žalik and B. Žalik 2011)	The separation measure of clusters is based on an overlap measure depending on the shape of the clusters.	

of the subsets in reference to the maximum diameter of every subset; which is defined as the maximum distance between a pair of points. Hence the Dunn index metric is defined as follows:

$$D(C) = \frac{C_C}{S_C} = \frac{min_{c_k \in C}\{min_{c_l \in C \setminus c_k} \delta(x_i, x_j)_{x_i \in c_k, x_j \in c_l}\}}{max_{c_k \in C} \Delta(c_k)} \tag{A.1}$$

where δ is the d_{inf} denoting the infimum function of the points in a cluster and Δ the diameter of a cluster.

An extension of the Dunn index is the *Generalised Dunn index* (Bezdek and Pal 1998) which address the sensitivity of changes in the cluster structure. More suitable definitions for the cluster diameter Δ and the set distance δ lead to validity indices for different types of clusters.

Calinski-Harabasz Index

Another important ratio-type cluster validation index is the *Calinski-Harabasz index* (Caliński and Harabasz 1974), which sums the squared distance of each point in a cluster from its centroid, to estimate the dispersion of the points in the cluster. The separation index is estimated in the same way but considering the centroids of all clusters. Hence, the equation that describes the Calinski-Harabasz index, is:

$$\begin{aligned} CH(C) &= \frac{\frac{S_C}{K-1}}{\frac{V_C}{n-K}} \Rightarrow CH(C) = \frac{n-K}{K-1} \cdot \frac{S_C}{V_C} \Rightarrow \\ CH(C) &= \frac{n-K}{K-1} \cdot \frac{\sum_{c_k \in C} \left(|c_k| \cdot ||\bar{c_k} - \bar{C}|| \right)}{\sum_{c_k \in C} \left(\sum_{x \in c_k} ||x - \bar{c_k}|| \right)} \end{aligned} \tag{A.2}$$

Davies-Bouldin Index

Similar to the Dunn index but based on the average error for each cluster is *Davies-Bouldin index* (Davies and Bouldin 1979), which identifies clusters which are far from each other and compact. Its estimation is based on the Minkowski metric of the centroids and the distance of the points in the cluster from its centroid. The function defining the similarity of two clusters is expressed below:

$$DB(C) = \frac{1}{|C|} \cdot \sum_{c_k \in C} max_{c_l \in C \setminus c_k} \frac{S(c_k) + S(c_l)}{||\bar{c_k} - \bar{c_l}||_p}$$

where, (A.3)

$$S(c_i) = \Big(\frac{1}{|c_i|} \cdot \sum_{x \in c_i} ||x - \bar{c_i}||_2^q\Big)^{1/q}$$

CS Index

CS index (Chou et al. 2004) is a ratio-type cluster validity index, ideal for clusters with different densities and/or sizes. The separation of the partition is estimated by the distance of the nearest neighbor and the cohesion by the diameter of the cluster. The CS index is defined as:

$$\begin{aligned} CS(C) &= \frac{C_C}{S_C} \Rightarrow \\ CS(C) &= \frac{\sum_{c_k \in C} \Big(\frac{1}{|c_k|} \cdot \sum_{x_i \in c_k} max_{x_j \in c_k} ||x_i - x_j||\Big)}{\sum_{c_k \in C} min_{c_l \in C \setminus c_k} ||\bar{c_k} - \bar{c_l}||} \end{aligned} \tag{A.4}$$

COP Index

Another ratio-type cluster index is the *COP index* (Gurrutxaga et al. 2010), which estimates the cohesion of the clusters by the distance of the points in a cluster to its centroid, and the separation by the distance of the furthest neighbor. The index is bounded between 0 and 1, and is defined as:

$$\begin{aligned} COP(C) &= \frac{1}{n} \cdot \sum_{c_k \in C} |c_k| \cdot \frac{C_{c_k}}{S_{c_k}} \Rightarrow \\ COP(C) &= \frac{1}{n} \cdot \sum_{c_k \in C} \Big(|c_k| \cdot \frac{\frac{1}{|c_k|} \cdot \sum_{x \in c_k} ||x - \bar{c_k}||}{min_{x_i \notin c_k} \big(max_{x_j \in c_k} ||x_i - x_j||\big)}\Big) \end{aligned} \tag{A.5}$$

SV-index

SV index (K. R. Žalik and B. Žalik 2011) is a ratio-type cluster validation index, for measuring the partition validity of clusters that differ in density and size. It estimates the separation of the clusters by considering the distance of the nearest neighbor and the cohesion by estimating

the distance of the points at the boarder of the cluster to its centroid.

$$
\begin{aligned}
SV(C) &= \frac{S_C}{V_C} \Rightarrow \\
SV(C) &= \frac{\sum_{c_k \in C} min_{c_l \in C \setminus c_k} ||\bar{c_k} - \bar{c_l}||}{\sum_{c_k \in C} \left(\frac{10}{|c_k|} \cdot \sum max_{x \in c_k} \big((0.1 \cdot |c_k|) \cdot ||x - \bar{c_k}|| \big) \right)}
\end{aligned}
\tag{A.6}
$$

OS-index

The cluster validity *OS index* (K. R. Žalik and B. Žalik 2011) is a ratio-type index based on the overlap and separation of a partition. Similar to the SV-index the separation is estimated by the distance of the nearest neighbor. However, different from the SV-index, the degree of overlap among all clusters is considered. The OS-index is expressed as:

$$
\begin{aligned}
OS(C) &= \frac{O_C}{S_C} \Rightarrow \\
OS(C) &= \frac{\sum_{c_k \in C} \sum_{x_i \in c_k} o(x_i, c_k)}{\sum_{c_k \in C} min_{c_l \in C \setminus c_k} ||\bar{c_k} - \bar{c_l}||}
\end{aligned}
$$

where,

$$
o(x_i, c_k) = \begin{cases} \frac{a(x_i, c_k)}{b(x_i, c_k)} & \text{if } \frac{b(x_i, c_k) - a(x_i, c_k)}{b(x_i, c_k) + a(x_i, c_k)} < 0.4 \\ 0 & \text{otherwise} \end{cases}
\tag{A.7}
$$

and

$$
\begin{aligned}
a(x_i, c_k) &= \frac{1}{|c_k|} \cdot \sum_{x_j \in c_k} ||x_i - x_j|| \\
b(x_i, c_k) &= \frac{1}{|c_k|} \cdot \sum_{x_j \notin c_k} min(|c_k| \cdot ||x_i - x_j||)
\end{aligned}
$$

The term $o(x_i, c_k)$ expresses the overlap degree of an object x_i belonging to more than one clusters c_k.

S_Dbw Index

The *S_Dbw index* (Halkidi and Vazirgiannis 2001) is a ratio-type cluster validation index and is based on the clusters' compactness and the density between clusters. The clusters' compactness is estimated by an average scattering for clusters, whereas the density estimation between the clusters considers the average density in the region among clusters compared with the density of the clusters. A lower density between clusters in comparison with the density of the clusters, indicates better separation of the clusters. The S_Dbw index is defined as:

$$
\begin{aligned}
&S\_Dbw(C) = V_C + D_C \Rightarrow \\
&S\_Dbw(C) = Scat(C) + Dens\_bw(C) \\
&Scat(C) = \frac{1}{K} \cdot \sum_{c_k \in C} \frac{||\sigma(c_k)||}{||\sigma(X)||} \\
&Dens\_bw = \frac{1}{K \cdot (K-1)} \cdot \sum_{c_k \in C} \sum_{c_l \in C \setminus c_k} \frac{density(c_k, c_l)}{max\{density(c_k), density(c_l)\}} \\
&\text{where,} \\
&density(c_k) = \sum_{x \in c_k} f(x, \bar{c_k}) \\
&density(c_k, c_l) = \sum_{x \in (c_k \cup c_l)} f\Big(x, \frac{\bar{c_k} + \bar{c_l}}{2}\Big) \\
&\text{and} \\
&f(x, \bar{c_k}) = \begin{cases} 0 & \text{if } ||x - \bar{c_k}|| > stdev(C) \\ 1 & \text{otherwise} \end{cases} \\
&stdev(C) = \frac{1}{K} \cdot \sqrt{\sum_{c_k \in C} ||\sigma(c_k)||}
\end{aligned}
\tag{A.8}
$$

The term $stdev(C)$ is the standard deviation of a partition C, where $||x|| = (x^T x)^{1/2}$ is the euclidean norm of a vector x. Moreover, the term $\sigma(X)$ is the variance of the dataset X and $\sigma(c_k)$ is the variance of the cluster c_k , which is equal to $\sigma(c_k) = \frac{1}{|c_k|} \cdot \sum_{x \in c_k} (x - \bar{c_k})^2$.

Silhouettes Index

The *Silhouettes index* (Rousseeuw 1987) is based on the average summation of the distances of the points in a cluster as well as the distance of the closest neighbour cluster. The equation

below describes the Silhouettes index:

$$
\begin{aligned}
&SI(C) = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{(b_i - a_i)}{max(a_i, b_i)} \\
&\text{where,} \\
&a(x, c_k) = \frac{1}{|c_k|} \cdot \sum_{x_i \in c_k} ||x_i - x|| \\
&b(x, c_k) = min_{c_l \in C/c_k} \Big(\frac{1}{|c_l|} \cdot \sum_{x_i \in c_l} ||x_i - x|| \Big)
\end{aligned}
\tag{A.9}
$$

Score Function Index

Score Function index (Saitta et al. 2007) performs best and is based on inter-cluster and intra-cluster distances. It combines the Between Class Distance (BCD) and the Within Class Distance (WCD) based on the comparison of centroids approach. The Score Function to maximise the BCD term, minimise the WCD term as well as be bounded, is described as follows:

$$
\begin{aligned}
&SF(C) = 1 - \frac{1}{e^{e^{bcd - wcd}}} \\
&\text{where,} \\
&bcd(C) = \frac{\sum_{c_k \in C} \big(||\bar{c_k} - \bar{C}|| \cdot |c_k| \big)}{|X| \cdot |C|} \\
&wcd(C) = \sum_{c_k \in C} \Big(\frac{1}{|c_k|} \cdot \sum_{x \in c_k} ||x - \bar{c_k}|| \Big)
\end{aligned}
\tag{A.10}
$$

Appendix B

Study of DiSR

B.1 Information & Instructions for Experiments

Figure B.1 shows the information given to the recruited participants for the study of DiSR. They were asked to read and understand them prior to filling in and signing the consent participation form.

Instructions for both the experimental tasks were also provided to the participants (Fig. B.2). They could ask for clarifications prior to starting each task. For task 1, the participants were asked to select only one of the relations that best describes the interaction happening in a scene. The dendrogram was provided to show the hierarchy of the selection/detection of relations. *E.g.*, in a scenario where an object is both contained in another object and is it being supported by the same object, due to the "contain" relation being higher in the hierarchy, the participant should appointing the "contain" relation for that specific interaction. Participants were also informed that a "support" relation needs a surface kind of object to hold, different from the "touching" relation.

UNIVERSITY OF LEEDS

Participant information sheet

Information sheet for participants
Depth-informed Spatial Relations Use

Before you decide whether to take part in this study it is important for you to understand why the research is being done and what it will involve. Please take your time to read the following information carefully and discuss it if you wish. Ask me if there is anything that is not clear or if you would like more information. Take your time to decide on whether or not you wish to take part.

What is the purpose of the study?
The aim of this study is to collect data about how people use the Depth-informed Spatial Relations (DiSR) to describe real world scenes and object interactions. Spatial relations are words or phrases that describe the spatial relationship between two of more objects, such as: 'touching', 'containing', 'supporting', etc..
In this study we will collect data on the two following tasks.

First task: The aim of this task is to evaluate how well the DiSR can be used to describe various scenes with interacting objects. For this task, you will be asked to give a set of DiSR relations along with the corresponding objects, that you believe best describes a given scene. You will be asked to repeat this for a number of scenes.

Second task: The goal of the second task is to evaluate how descriptive the DiSRs are. Given a set of DiSRs, you will be asked to create the real world representation of object interactions that can be best described with the given set of relations. You will be asked to repeat the task for several sets of DiSRs.

What will happen to me if I take part?
After being introduced to the experiment (reading consent form, instructions, etc..) you will be asked to spend 20 minutes at the computer completing a task which involves some or all of the following: navigating a virtual 3D environment using the keyboard and mouse, selecting objects with the mouse and providing text descriptions of objects in the environment.

Do I have to take part?
It is completely up to you whether to take part or not. If you consent to taking part then you are free to withdraw from the research within **two weeks** without having to give a reason. Participation in this study is voluntary and if you have any questions please discuss these with one of the research team. If you do decide to take part you will be given this information sheet to keep (and be asked to sign a consent form).

Will my taking part in the study be kept confidential?/What will happen to the results of the research project
All data will be kept strictly confidential and any individual data in write-ups/publications will be referred to by code-name only. All of the data obtained from the participants during the study will be anonymised. Data placed on computers for analyses will be anonymised. Only the investigators and research assistants will have access to these files.

Contact for further information
If you require any further information about the study and its results or have any questions and/or worries, please contact Alexia Toumpa (scat@leeds.ac.uk).

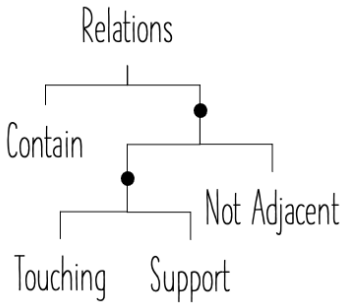
Figure B.1: Information given to the participants.

Spatial Relations Study - Instructions

This study involves the set of DiSR relations, which are described below:

Contain(*x*,*y*)	object *x* is containing object *y*
Touching(*x*,*y*)	object *x* is touching object *y*
Support(*x*,*y*)	object *x* supports object *y*
NotAdjacent(*x*,*y*)	object *x* is not spatially touching object *y*

DiSR Detection Hierarchy



In this study you will need to complete the following two tasks:

Task 1:
The aim of this task is to evaluate how well the DiSR can be used to describe various scenes with interacting objects. In this task you will be given 16 scenes with a pair of interacting objects each (visualized in a 3D virtual environment), and you will be asked to select one of the DiSR relations that best describes the interaction taking place, in each scene.

Task 2:
The goal of the second task is to evaluate how descriptive the DiSR relations are. You will be given 16 DiSR relations holding between a pair of objects each, and you will be asked to create the real-world representation of the specific's pair interactions that you believe describes best each relation. For this task, you will use a 3D virtual environment, in which you will be able to move objects around and resize them.

Move object:
Select the arrow pointing towards the direction you want the object to be moved at, and drag the object towards that direction.
Resize object:
Select the cube on the axis you want the object to be resized and drag it along that axis.
Rotate object:
Select the arc along which you want the object to be rotated, and drag your pointer along the arc.

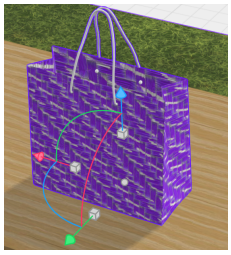


Figure B.2: Instructions given to the participants.

Appendix C

Qualitative Results of Video Frame Prediction

C.1 Results of FP-LSTM

Further qualitative results of FP-LSTM are presented for the Moving MNIST dataset in Figure C.1 and Figure C.2, for the UCF101 dataset in Figure C.3 and Figure C.4, and for the CAD-120 dataset in Figure C.5, Figure C.6, Figure C.7, Figure C.8, and Figure C.9.

C.2 Results from CPL

Further qualitative results of CPL are presented for the CAD-120 dataset in Figure C.10 and Figure C.11.

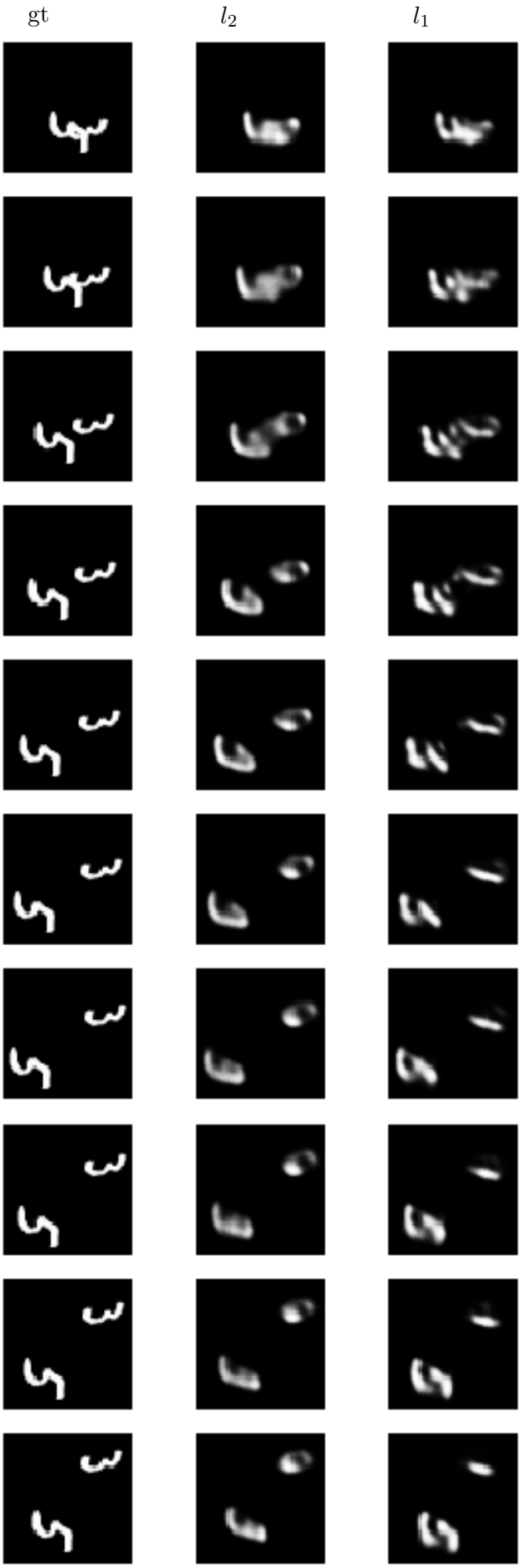


Figure C.1: FP-LSTM predictions on the Moving MNIST dataset for moving digits 3 and 5. The model was trained with a single FP-LSTM layer (l_1) and a double layer (l_2) separately. Ground truth image frames are shown in the "gt" rows.

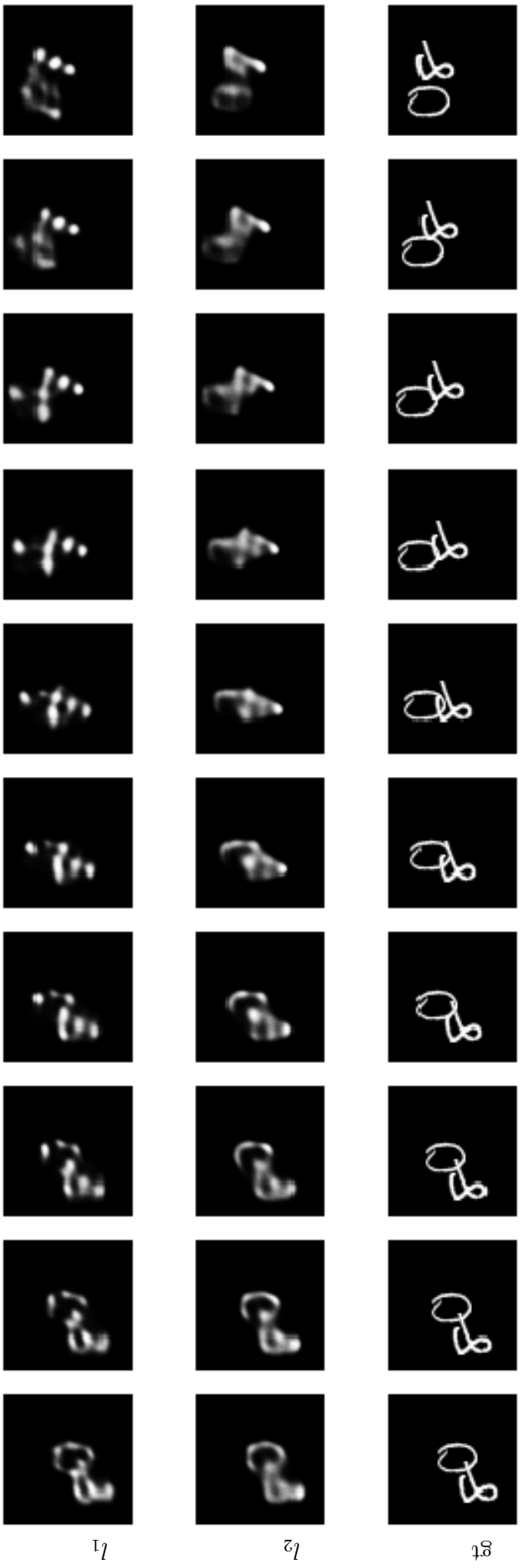


Figure C.2: FP-LSTM predictions on the Moving MNIST dataset for moving digits 8 and 0. The model was trained with a single FP-LSTM layer (l_1) and a double layer (l_2) separately. Ground truth image frames are shown in the "gt" rows.

Figure C.3: FP-LSTM predictions on the UCF101 dataset for a scene without a lot of spatio-temporal variations between the visualized frames of the video sequence. The model was trained with a single FP-LSTM layer (l_1) and a double layer (l_2) separately. Ground truth image frames are shown in the "gt" rows.

Figure C.4: FP-LSTM predictions on the UCF101 dataset with some spatio-temporal variation between the frames in the presented video sequence. The model was trained with a single FP-LSTM layer (l_1) and a double layer (l_2) separately. Ground truth image frames are shown in the "gt" rows.

Figure C.5: FP-LSTM predictions on the "microwaving food"activity of the CAD-120 dataset. The model was trained with a single FP-LSTM layer (l_1) and a double layer (l_2) separately. Ground truth image frames are shown in the "gt" rows.

Figure C.6: FP-LSTM predictions on the "cleaning objects" activity of the CAD-120 dataset. The model was trained with a single FP-LSTM layer (l_1) and a double layer (l_2) separately. Ground truth image frames are shown in the "gt" rows.

Figure C.7: FP-LSTM predictions on the "making cereal"activity of the CAD-120 dataset. The model was trained with a single FP-LSTM layer (l_1) and a double layer (l_2) separately. Ground truth image frames are shown in the "gt" rows.

Figure C.8: FP-LSTM predictions on the "taking food" activity of the CAD-120 dataset. The model was trained with a single FP-LSTM layer (l_1) and a double layer (l_2) separately. Ground truth image frames are shown in the "gt" rows.

Figure C.9: FP-LSTM predictions on the "cleaning objects" activity of the CAD-120 dataset. The model was trained with a single FP-LSTM layer (l_1) and a double layer (l_2) separately. Ground truth image frames are shown in the "gt" rows.

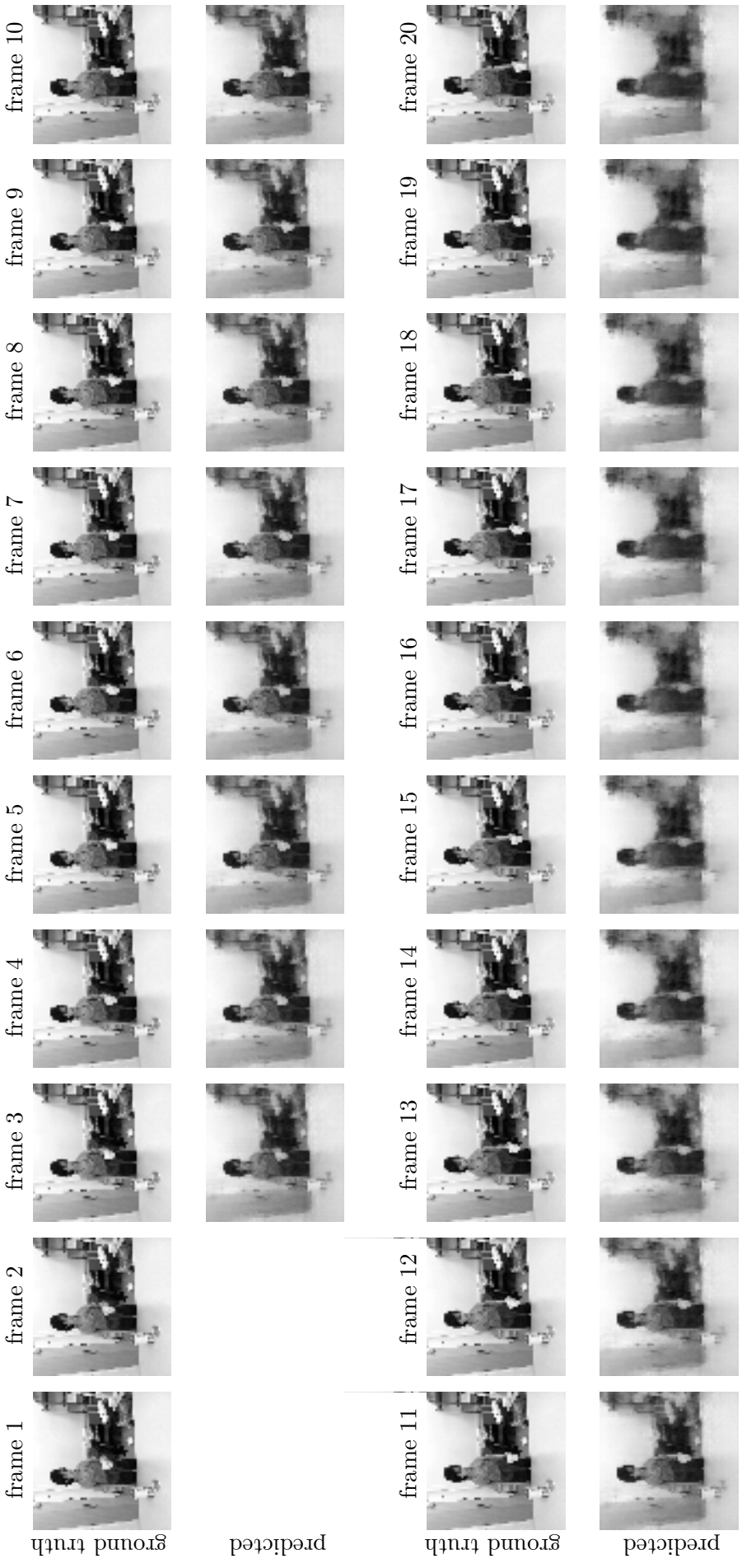


Figure C.10: CPL predictions for "taking medicine" video activity of the CAD-120 dataset.

Figure C.11: CPL predictions for "taking food" video activity of the CAD-120 dataset.

References

Aksoy, Eren Erdal, Abramov, Alexey, Dörr, Johannes, Ning, Kejun, Dellen, Babette, and Wörgötter, Florentin (2011). "Learning the semantics of object–action relations by observation". In: *The International Journal of Robotics Research* 30.10, pp. 1229–1249.

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

Allen, James F (1990). "Maintaining Knowledge about Temporal Intervals". In: *Readings in Qualitative Reasoning about Physical Systems*. Elsevier, pp. 361–372.

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

Arbelaitz, Olatz, Gurrutxaga, Ibai, Muguerza, Javier, PéRez, JesúS M, and Perona, IñIgo (2013). "An Extensive Comparative Study of Cluster Validity Indices". In: *Pattern Recognition* 46.1, pp. 243–256.

Bennett, Brandon, Chaudhri, Vinay, and Dinesh, Nikhil (2013). "A vocabulary of topological and containment relations for a practical biological ontology". In: *Spatial Information Theory: 11th International Conference, COSIT 2013, Scarborough, UK, September 2-6, 2013. Proceedings 11*. Springer, pp. 418–437.

Berry, Michael JA and Linoff, Gordon S (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

Bezdek, James C and Pal, Nikhil R (1998). "Some New Indexes of Cluster Validity". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 28.3, pp. 301–315.

Bolme, David S, Beveridge, J Ross, Draper, Bruce A, and Lui, Yui Man (2010). "Visual Object Tracking using Adaptive Correlation Filters". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 2544–2550.

Byeon, Wonmin, Wang, Qin, Kumar Srivastava, Rupesh, and Koumoutsakos, Petros (2018). "Contextvp: Fully context-aware video prediction". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 753–769.

Caliński, Tadeusz and Harabasz, Jerzy (1974). "A Dendrite Method for Cluster Analysis". In: *Communications in Statistics-theory and Methods* 3.1, pp. 1–27.

Chen, Geng, Zhang, Wendong, Lu, Han, Gao, Siyu, Wang, Yunbo, Long, Mingsheng, and Yang, Xiaokang (2022). "Continual Predictive Learning from Videos". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10728–10737.

Chen, Juan, Cohn, Anthony G, Liu, Dayou, Wang, Shengsheng, Ouyang, Jihong, and Yu, Qiangyuan (2015). "A Survey of Qualitative Spatial Representations". In: *The Knowledge Engineering Review* 30.1, pp. 106–136.

Chou, C-H, Su, M-C, and Lai, Eugene (2004). "A New Cluster Validity Measure and its Application to Image Compression". In: *Pattern Analysis and Applications* 7.2, pp. 205–220.

Chuang, Ching-Yao, Li, Jiaman, Torralba, Antonio, and Fidler, Sanja (2018). "Learning to Act Properly: Predicting and Explaining Affordances from Images". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 975–983.

Cohn, Anthony G (1995). "A hierarchical representation of qualitative shape based on connection and convexity". In: *International Conference on Spatial Information Theory*. Springer, pp. 311–326.

Cohn, Anthony G, Bennett, Brandon, Gooday, John, and Gotts, Nicholas Mark (1997). "Qualitative Spatial Representation and Reasoning with the Region Connection Calculus". In: *GeoInformatica* 1.3, pp. 275–316.

Damen, Dima, Doughty, Hazel, Farinella, Giovanni Maria, Fidler, Sanja, Furnari, Antonino, Kazakos, Evangelos, Moltisanti, Davide, Munro, Jonathan, Perrett, Toby, Price, Will, and Wray, Michael (2018). "Scaling Egocentric Vision: The EPIC-KITCHENS Dataset". In: *European Conference on Computer Vision (ECCV)*.

– (2021). "The EPIC-KITCHENS Dataset: Collection, Challenges and Baselines". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 43.11, pp. 4125–4141. DOI: 10.1109/TPAMI.2020.2991965.

Damen, Dima, Doughty, Hazel, Farinella, Giovanni Maria, Furnari, Antonino, Ma, Jian, Kazakos, Evangelos, Moltisanti, Davide, Munro, Jonathan, Perrett, Toby, Price, Will, and Wray, Michael (2022). "Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100". In: *International Journal of Computer Vision (IJCV)* 130, pp. 33–55. URL: https://doi.org/10.1007/s11263-021-01531-2.

Davies, David L and Bouldin, Donald W (1979). "A Cluster Separation Measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2, pp. 224–227.

Delafontaine, Matthias, Cohn, Anthony G, and Van de Weghe, Nico (2011). "Implementing a Qualitative Calculus to Analyse Moving Point Objects". In: *Expert Systems with Applications* 38.5, pp. 5187–5196.

Deng, Shengheng, Xu, Xun, Wu, Chaozheng, Chen, Ke, and Jia, Kui (2021). "3D AffordanceNet: A Benchmark for Visual Object Affordance Understanding". In: *arXiv preprint arXiv:2103.16397*.

Denton, Emily L et al. (2017). "Unsupervised Learning of Disentangled Rrepresentations from Video". In: *Advances in Neural Information Processing Systems*, pp. 4414–4423.

Do, Thanh-Toan, Nguyen, Anh, and Reid, Ian (2018). "AffordanceNet: An End-to-End Deep Learning Approach for Object Affordance Detection". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 1–5.

Dunn, Joseph C (1973). "A Fuzzy Relative of the ISODATA Process and its use in Detecting Compact Well-Separated Clusters". In: *J. Cybernet* 3.3, pp. 32–57.

Dylla, Frank, Lee, Jae Hee, Mossakowski, Till, Schneider, Thomas, Delden, André Van, Ven, Jasper Van De, and Wolter, Diedrich (2017). "A Survey of Qualitative Spatial and Temporal Calculi: Algebraic and Computational Properties". In: *ACM Computing Surveys (CSUR)* 50.1, pp. 1–39.

Egenhofer, Max J (2005). "Spherical topological relations". In: *Journal on Data Semantics III*. Springer, pp. 25–49.

Egenhofer, Max J, Clementini, Eliseo, and Di Felice, Paolino (1994). "Topological relations between regions with holes". In: *International Journal of Geographical Information Science* 8.2, pp. 129–142.

Egenhofer, Max J and Franzosa, Robert D (1991). "Point-set topological spatial relations". In: *International Journal of Geographical Information System* 5.2, pp. 161–174.

– (1995). "On the equivalence of topological relations". In: *International Journal of Geographical Information Systems* 9.2, pp. 133–152.

Egenhofer, Max J and Herring, John (1990). "Categorizing binary topological relations between regions, lines, and points in geographic databases". In: *The* 9.94-1, p. 76.

Egenhofer, Max J, Mark, David M, and Herring, John (1994). "The 9-Intersection: Formalism and Its Use for Natural-Language Spatial Predicates". In: *Report*. Vol. 9, p. 76.

Egenhofer, Max J and Sharma, Jayant (1993). "Topological relations between regions in ρ^2 and $\mathbb{Z}^2$". In: *International Symposium on Spatial Databases*. Springer, pp. 316–336.

Egenhofer, Max J and Vasardani, Maria (2007). "Spatial reasoning with a hole". In: *Spatial Information Theory: 8th International Conference, COSIT 2007, Melbourne, Australiia, September 19-23, 2007. Proceedings 8*. Springer, pp. 303–320.

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

Finn, Chelsea, Goodfellow, Ian, and Levine, Sergey (2016). "Unsupervised Learning for Physical Interaction Through Video Prediction". In: *Advances in Neural Information Processing Systems*, pp. 64–72.

Frank, Andrew U (1991). "Qualitative spatial reasoning with cardinal directions". In: *7. Österreichische Artificial-Intelligence-Tagung/Seventh Austrian Conference on Artificial Intelligence: Wien, Austria, 24.–27. September 1991 Proceedings*. Springer, pp. 157–167.

Freksa, Christian (1992). "Temporal reasoning based on semi-intervals". In: *Artificial Intelligence* 54.1-2, pp. 199–227.

Galton, Antony (1994). "Lines of Sight". In: *AISB Workshop on Spatial and Spatio-Temporal Reasoning*. Vol. 35, pp. 37–39.

Galton, Antony and Meathrel, Richard C (1999). "Qualitative outline theory". In: *IJCAI*. Citeseer, pp. 1061–1066.

Gatsoulis, Yiannis, Alomari, Muhannad, Burbridge, Chris, Dondrup, Christian, Duckworth, Paul, Lightbody, Peter, Hanheide, Marc, Hawes, Nick, Hogg, David C, and Cohn, Anthony G (2016). "QSRlib: A Software Library for Online Acquisition of Qualitative Spatial Relations from Video". In: *In Workshop on Qualitative Reasoning (QR16), at IJCAI*.

Gerevini, Alfonso and Renz, Jochen (2002). "Combining Topological and Size Information for Spatial Reasoning". In: *Artificial Intelligence* 137.1-2, pp. 1–42.

Gibson, James J (1977). "The Theory of Affordances". In: *Hilldale, USA* 1.2.

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

Gooday, JM and Cohn, AG (1994). "Conceptual neighborhoods in temporal and spatial reasoning". In: *Spatial and Temporal Reasoning, ECAI* 94.

Gottfried, Björn (2003a). "Tripartite line tracks qualitative curvature information". In: *International Conference on Spatial Information Theory*. Springer, pp. 101–117.

Gottfried, Björn (2003b). "Tripartite line tracks–bipartite line tracks". In: *Annual Conference on Artificial Intelligence*. Springer, pp. 535–549.

– (2004). "Reasoning About Intervals in Two Dimensions". In: *Systems, Man and Cybernetics, 2004 IEEE International Conference on*. Vol. 6. IEEE, pp. 5324–5332.

Gurrutxaga, Ibai, Albisua, Iñaki, Arbelaitz, Olatz, Martín, José I, Muguerza, Javier, Pérez, Jesús M, and Perona, Iñigo (2010). "SEP/COP: An Efficient Method to Find the Best Partition in Hierarchical Clustering based on a new Cluster Validity Index". In: *Pattern Recognition* 43.10, pp. 3364–3373.

Halkidi, Maria and Vazirgiannis, Michalis (2001). "Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set". In: *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. IEEE, pp. 187–194.

He, Kaiming, Gkioxari, Georgia, Dollár, Piotr, and Girshick, Ross (2017). "Mask R-CNN". In: *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, pp. 2980–2988.

Hou, Zhi, Yu, Baosheng, Qiao, Yu, Peng, Xiaojiang, and Tao, Dacheng (2021). "Affordance Transfer Learning for Human-Object Interaction Detection". In: *arXiv preprint arXiv:2104.02867*.

Isli, Amar, Haarslev, Volker, Möller, Ralf, et al. (2001). *Combining cardinal direction relations and relative orientation relations in qualitative spatial reasoning*. Univ., Bibliothek des Fachbereichs Informatik.

Kalchbrenner, Nal, Oord, Aäron, Simonyan, Karen, Danihelka, Ivo, Vinyals, Oriol, Graves, Alex, and Kavukcuoglu, Koray (2017). "Video Pixel Networks". In: *International Conference on Machine Learning*. PMLR, pp. 1771–1779.

Kim, Minho and Ramakrishna, RS (2005). "New Indices for Cluster Validity Assessment". In: *Pattern Recognition Letters* 26.15, pp. 2353–2363.

Kjellström, Hedvig, Romero, Javier, and Kragić, Danica (2011). "Visual object-action recognition: Inferring object affordances from human demonstration". In: *Computer Vision and Image Understanding* 115.1, pp. 81–90.

Köhler, Christian (2002). "The Occlusion Calculus". In: *Cognitive Vision Workshop*. Citeseer, pp. 420–450.

Kokic, Mia, Stork, Johannes A, Haustein, Joshua A, and Kragic, Danica (2017). "Affordance Detection for Task-Specific Grasping using Deep Learning". In: *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*. IEEE, pp. 91–98.

Koppula, Hema Swetha, Gupta, Rudhir, and Saxena, Ashutosh (2013). "Learning Human Activities and Object Affordances from RGB-D videos". In: *The International Journal of Robotics Research* 32.8, pp. 951–970.

Landsiedel, Christian, Rieser, Verena, Walter, Matthew, and Wollherr, Dirk (2017). "A review of spatial reasoning and interaction for real-world robotics". In: *Advanced Robotics* 31.5, pp. 222–242.

LeCun, Yann A, Bottou, Léon, Orr, Genevieve B, and Müller, Klaus-Robert (2012). "Efficient Backprop". In: *Neural networks: Tricks of the trade*. Springer, pp. 9–48.

Leyton, Michael (1988). "A Process-Grammar for Shape". In: *Artificial Intelligence* 34.2, pp. 213–247.

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

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

Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, and Zitnick, C Lawrence (2014). "Microsoft COCO: Common Objects in Context". In: *European Conference on Computer Vision*. Springer, pp. 740–755.

Lotter, William, Kreiman, Gabriel, and Cox, David (2016). "Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning". In: *arXiv preprint arXiv:1605.08104*.

Moldovan, Bogdan and De Raedt, Luc (2014). "Occluded Object Search by Relational Affordances". In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 169–174.

Montesano, Luis and Lopes, Manuel (2009). "Learning Grasping Affordances from Local Visual Descriptors". In: *2009 IEEE 8th International Conference on Development and Learning*. IEEE, pp. 1–6.

Moratz, Reinhard (2006). "Representing relative direction as a binary relation of oriented points". In: *ECAI*. Vol. 6, pp. 407–411.

Moratz, Reinhard and Ragni, Marco (2008). "Qualitative spatial reasoning about relative point position". In: *Journal of Visual Languages & Computing* 19.1, pp. 75–98.

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

Narayanan, Annamalai, Chandramohan, Mahinthan, Venkatesan, Rajasekar, Chen, Lihui, Liu, Yang, and Jaiswal, Shantanu (2017). "graph2vec: Learning Distributed Representations of Graphs". In: *arXiv preprint arXiv:1707.05005*.

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

– (2017). "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)*. IEEE, pp. 5908–5915.

Oreifej, Omar and Liu, Zicheng (2013). "HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 716–723.

Pal, Nikhil R and Biswas, J (1997). "Cluster Validation using Graph Theoretic Concepts". In: *Pattern Recognition* 30.6, pp. 847–857.

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

– (2014). "Recognizing Object Affordances in Terms of Spatio-Temporal Object-Object Relationships". In: *International Conference on Humanoid Robots, November 18-20th 2014, Madrid, Spain*. IEEE conference proceedings, pp. 52–58.

Qi, Siyuan, Huang, Siyuan, Wei, Ping, and Zhu, Song-Chun (2017). "Predicting Human Activities Using Stochastic Grammar". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1164–1172.

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

Randell, David, Witkowski, Mark, and Shanahan, Murray (2001). "From Images to Bodies: Modelling and Exploiting Spatial Occlusion and Motion Parallax". In: *IJCAI*, pp. 57–66.

Randell, David A, Cui, Zhan, and Cohn, Anthony G (1992). "A Spatial Logic based on Regions and Connection". In: *KR* 92, pp. 165–176.

Rosenberg, Andrew and Hirschberg, Julia (2007). "V-measure: A conditional entropy-based external cluster evaluation measure". In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 410–420.

Rousseeuw, Peter J (1987). "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis". In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65.

Saitta, Sandro, Raphael, Benny, and Smith, Ian FC (2007). "A Bounded Index for Cluster Validity". In: *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, pp. 174–187.

Santos, Paulo E (2007). "Reasoning about depth and motion from an observer's viewpoint". In: *Spatial Cognition & Computation* 7.2, pp. 133–178.

Sawatzky, Johann, Srikantha, Abhilash, and Gall, Juergen (2017). "Weakly Supervised Affordance Detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2795–2804.

Schuldt, Christian, Laptev, Ivan, and Caputo, Barbara (2004). "Recognizing Human Actions: A Local SVM Approach". In: *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. Vol. 3. IEEE, pp. 32–36.

Shervashidze, Nino, Schweitzer, Pascal, Van Leeuwen, Erik Jan, Mehlhorn, Kurt, and Borgwardt, Karsten M (2011). "Weisfeiler-lehman graph kernels". In: *Journal of Machine Learning Research* 12.9.

Shi, Xingjian, Chen, Zhourong, Wang, Hao, Yeung, Dit-Yan, Wong, Wai-Kin, and Woo, Wang-chun (2015). "Convolutional LSTM network: A machine learning approach for precipitation nowcasting". In: *Advances in Neural Information Processing Systems* 28.

Siskind, Jeffrey Mark (1994). "Grounding language in perception". In: *Artificial Intelligence Review* 8, pp. 371–391.

– (2001). "Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic". In: *Journal of artificial intelligence research* 15, pp. 31–90.

Skiadopoulos, Spiros and Koubarakis, Manolis (2004). "Composing cardinal direction relations". In: *Artificial Intelligence* 152.2, pp. 143–171.

– (2005). "On the consistency of cardinal direction constraints". In: *Artificial Intelligence* 163.1, pp. 91–135.

Soomro, Khurram, Zamir, Amir Roshan, and Shah, Mubarak (2012). "UCF101: A dataset of 101 Human Actions Classes from Videos in the wild". In: *arXiv preprint arXiv:1212.0402*.

Sridhar, Muralikrishna, Cohn, Anthony G, and Hogg, David C (2008). "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)*. IOS Press, pp. 606–610.

– (2010a). “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.

– (2010b). “Unsupervised Learning of Event Classes from Video”. In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. AAAI Press, pp. 1631–1638.

Srivastava, Nitish, Mansimov, Elman, and Salakhudinov, Ruslan (2015). “Unsupervised Learning of Video Representations using LSTMs”. In: *International Conference on Machine Learning*. PMLR, pp. 843–852.

Sung, Jaeyong, Ponce, Colin, Selman, Bart, and Saxena, Ashutosh (2012). “Unstructured Human Activity Detection from RGBD images”. In: *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, pp. 842–849.

Tan, Haoliang, 0003, Le Wang, Zhang, Qilin, Gao, Zhanning, Zheng, Nanning, and Hua, Gang (2019). “Object Affordances Graph Network for Action Recognition”. In: *BMVC*, p. 145.

Toumpa, Alexia and Cohn, Anthony G (2019). “Relational Graph Representation Learning for Predicting Object Affordances”. In: *NeurIPS Workshop on Graph Representation Learning*.

– (2020). “Depth-informed Qualitative Spatial Representations for Object Affordance Prediction”. In: *33rd International Workshop on Qualitative Reasoning*.

– (2023a). “Future Qualitative Activity Graph Prediction”. In: *37th AAAI Conference on Artificial Intelligence, 3rd Workshop on Graphs and more Complex Structures for Learning and Reasoning (GCLR)*.

– (2023b). “Object-agnostic Affordance Categorization via Unsupervised Learning of Graph Embeddings”. In: *Journal of Artificial Intelligence Research* 77, pp. 1–38.

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

Van de Weghe, Nico, Cohn, Anthony G, and De Maeyer, Philippe (2004). “A Qualitative Representation of Trajectory Pairs”. In: *ECAI*. Vol. 16, p. 1103.

Van de Weghe, Nico, Cohn, Anthony G, De Tre, Guy, and De Maeyer, Philippe (2006). "A Qualitative Trajectory Calculus as a Basis for Representing Mmoving Objects in Geographical Information Systems". In: *Control and Cybernetics* 35.1, pp. 97–119.

Van Rijsbergen, C (1979). "Information retrieval: theory and practice". In: *Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems*. Vol. 79.

Villegas, Ruben, Yang, Jimei, Hong, Seunghoon, Lin, Xunyu, and Lee, Honglak (2017). "Decomposing Motion and Content for Natural Video Ssequence Prediction". In: *arXiv preprint arXiv:1706.08033*.

Wang, Aria Yuan and Tarr, Michael J (2020). "Learning Intermediate Features of Object Affordances with a Convolutional Neural Network". In: *arXiv preprint arXiv:2002.08975*.

Wang, Jiang, Liu, Zicheng, Wu, Ying, and Yuan, Junsong (2012). "Mining actionlet ensemble for action recognition with depth cameras". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1290–1297.

Wang, Yunbo, Jiang, Lu, Yang, Ming-Hsuan, Li, Li-Jia, Long, Mingsheng, and Fei-Fei, Li (2018). "Eidetic 3D LSTM: A Model for Video Prediction and Beyond". In: *International Conference on Learning Representations*.

Wei, Shih-En, Ramakrishna, Varun, Kanade, Takeo, and Sheikh, Yaser (2016). "Convolutional Pose Machines". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724–4732.

Wu, Chenxia, Zhang, Jiemi, Savarese, Silvio, and Saxena, Ashutosh (2015). "Watch-n-Patch: Unsupervised Understanding of Actions and Relations". In: *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, pp. 4362–4370.

Wu, Hongtao, Misra, Deven, and Chirikjian, Gregory S (2020). "Is That a Chair? Imagining Affordances Using Simulations of an Articulated Human Body". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 7240–7246.

Xu, Jingwei, Ni, Bingbing, Li, Zefan, Cheng, Shuo, and Yang, Xiaokang (2018). "Structure Preserving Video Prediction". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1460–1469.

Xu, Ruinian, Chu, Fu-Jen, Tang, Chao, Liu, Weiyu, and Vela, Patricio A (2021). "An Affordance Keypoint Detection Network for Robot Manipulation". In: *IEEE Robotics and Automation Letters* 6.2, pp. 2870–2877.

Yanardag, Pinar and Vishwanathan, SVN (2015). "Deep graph kernels". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1365–1374.

Yao, Bangpeng, Ma, Jiayuan, and Fei-Fei, Li (2013). "Discovering Object Functionality". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2512–2519.

Žalik, Krista Rizman and Žalik, Borut (2011). "Validity Index for Clusters of Different Sizes and Densities". In: *Pattern Recognition Letters* 32.2, pp. 221–234.

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

Zhao, Ying and Karypis, George (2001). "Criterion Functions for Document Clustering: Experiments and Analysis". In: pp. 1–40.