Relational Graph Representation Learning for Predicting Object Affordances

Alexia Toumpa and Anthony G. Cohn
School of Computing
University of Leeds
{A.Toumpa, A.G.Cohn}@leeds.ac.uk

Abstract

We address the problem of affordance classification for class-agnostic objects considering an open set of actions, by unsupervised learning of object interactions, inducing object affordance classes. A novel qualitative spatial representation incorporating depth information is used to construct *Activity Graphs* which encode object interactions. These Activity Graphs are clustered to obtain interaction classes, and subsequently extract classes of object affordances. Our experiments demonstrate that our method learns object affordances without being scene- or object-specific.

1 Introduction

In the literature, the term *affordance* of an object differs depending on the context. In some 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, *e.g.* the hand of a hammer has the affordance of 'hold' whereas the head has the affordance of 'hit'. In contrast, when considering a human-object interaction recognition task, *affordance* can be interpreted as the way an object is afforded by the human in a scene, *e.g.* if a human uses a book as a hammer then the book will have the affordance of 'hit'. Moreover, any object may have more than one affordance as it depends on the purpose it is being used for, *e.g.* a book can have the affordance of 'hit' when it is being utilized as a hammer or 'hold' when it plays the role of a tray, and such multi-labelled affording objects can be recognized by considering their interactions with other objects [1].

Several methods have been proposed for detecting functional object parts as well as affordance labels. Some works involve the detection of object affordance parts by considering their visual characteristics and their geometric features [2, 3, 4, 5, 6, 7]. Incorporating knowledge of the scene and context in which an object is being used boosts prediction accuracy [8, 9]. However, the task becomes challenging when no scene or object restrictions take place and the affordance space enlarges. For this purpose, many works have considered exploiting the correlation of human actions and the detected objects in a scene for predicting object affordances, *e.g.* a DL architecture for inferring human-object interactions [10], as well as the people's skeletal data to predict the functionality of the detected objects [11]. Fang et al. [12] exploit demonstration video data for reasoning about affordances by predicting the affordance location on the object along with the action of a human agent. Human-object and object-object interactions are also considered by exploiting object trajectories [13, 14].

However, object occlusion is one of the fundamental obstacles in these works. To overcome this, some research has focused on tracking the occlusion and detecting the containers and the containee objects in a scene [15, 16, 17, 18] assuming a predefined set of interactions/affordances. Nonetheless, such an assumption restricts the method to be applied to a limited number of domains. Another stream of works introduces graph structures of object interactions to accommodate domain independence, *e.g.* a graph representation captures qualitative spatio-temporal relationships between objects and thus infers event classes from which a hierarchy of functional object categories is induced [19]; occlusions are not explicitly handled though. Aksoy et al. [20] also employ a graph representation

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

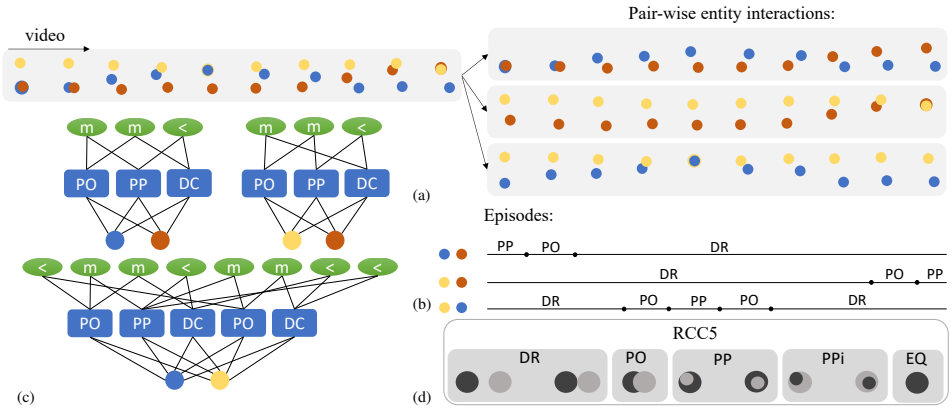


Figure 1: AsG construction from pair-wise object interactions captured in a video scene.

based on the scene semantics, correlating segmented objects in reference to their interactions, though a pre-specified set of objects is solely considered with a very limited set of affordances.

In this work we present a novel approach for addressing the problem of affordance classification by exploiting pair-wise object interactions in RGB-D data. By learning a high-level representation of interactions, our approach is not limited to any specific scene type, or a predefined number of objects, and it tackles the issue of temporally occluded objects. The proposed methodology is orthogonal to the employed set of affordances, and therefore can be extended to incorporate any affordance provided the availability of visual data. To abstract from the continuous space of spatio-temporal interactions, we employ a high-level graph structure, the *Activity Graph*, for representing pair-wise object interactions while acquiring novel depth-enhanced qualitative spatio-temporal relations between pairs of objects to handle occlusions. Our methodology examines object affordances in the context of interactions, thus the affordance of an object is characterized by the interaction it is involved in. Affordances of objects are deduced from their pair-wise interactions, in reference to their Activity Graphs. From these graph representations we learn in an unsupervised way affordance classes through exploiting their intra-class graph similarity and clustering them into groups. Clustering graph structures produces a hierarchical tree representation which demonstrates their graph similarity.

2 Methodology

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. From a video demonstrating a human activity, we represent the spatio-temporal relational changes, describing interactions of objects in the scene, by employing a graph structure called an *Activity Graph*, discussed further in this section. A clustering mechanism, is also introduced, acting on these graphs to produce classes of affordances in reference to their *Activity Graph* similarities.

Activity Graph. Sequences of interactions which are combined to obtain an activity can be deduced from an *Activity Graph*. An Activity Graph (AG) [21], is a graph representation which captures spatio-temporal information of the interactions between entities present in a video sequence. Let $G = (V, E)$ be an AG, where the vertices V are partitioned into 3 layers $V_{entities}$, V_{spat}, and V_{temp} and the edges E exists only between adjacent layers to represent pair-wise entity interactions. Each set of vertices carries a different kind of information about the activity performed. $V_{entities}$ is the set of entities which interact, V_{spat} corresponds to the set of spatial relations which occur during an activity, and V_{temp} holds the temporal information of the occurrence of the spatial relations.

Initially, the set V_{spat} is determined by the objects' interactions in the 2D image plane and corresponds to the spatial interactions of their bounding boxes. To abstract from continuous space we capture qualitative relations to acquire more abstract representations. More specifically, we exploit the Region Connection Calculus (RCC5) relations [22, 23] as illustrated in Fig. 1(d). V_{temp} encodes the temporal relationships between the episodes over which particular spatial relationships hold. We exploit *Allen's interval algebra* [24] which consist of the relations: 'before' ($<$, $>$), 'meets' (m, mi), 'overlaps', (o, oi), 'starts' (s, si), 'during' (d, di), 'finishes' (f, fi), and 'equals' ($=$).

An object interaction is captured from a subgraph of an AG, called *Activity sub-Graph* or AsG, which comprise the subset V^s_{spat} of V_{spat} and V^s_{temp} of V_{temp} vertices indicating a particular interaction between a corresponding pair of objects, $V^s_{entities}$. Such AsGs are depicted in Fig. 1(c).

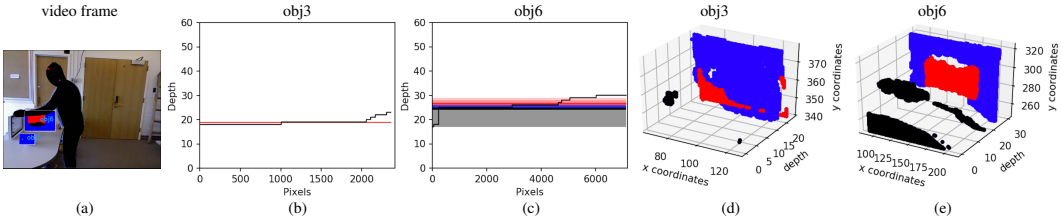


Figure 2: (a)frame with detected objects. Depth distribution of convex (b,d) and concave (c,e) objects.

The sequence of spatial relations obtained in every pair-wise interaction is extracted from the presence of *episodes*, which represent 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. Fig. 1(b) shows the episodes which are extracted from all the pair-wise entity interactions in a video.

Depth-enhanced Spatial Relations. Though simple and discrete object interactions are captured efficiently by employing the RCC5 relations in the image plane, the determination of more complex spatial relational structures is challenging, especially when considering a cluttered scene, i.e. occlusion of objects prevent detection of an interaction. To address this limitation we exploit the depth information of RGB-D video data in order to infer more accurately the objects' relative positions.

Without loss of generality, we focus on creating groups of the affordances *contain* and *support* as these are the most prominent and their distinction is challenging in view of object interactions and occlusions. The inference of these object affordances relies only on basic types of interactions; however due to the occlusion of objects without the utilization of depth information their differentiation is not a simple task. Though RCC5 is dimension independent, we enhance the calculation of RCC5 relations with the depth information available of the involved objects, rather than only exploiting their 2D projections in the camera plane, acquiring knowledge about their convexity-type. These enhanced relations are employed for producing the V_{spat} set of the AGs.

We introduce depth-enhanced RCC5 (D-RCC5) spatial relations aiming at grouping the objects in the scene into three convexity-based categories: 'convex', 'concave', and 'surface', in reference to their depth distribution, as depicted in Fig. 2. By estimating the distribution of the depth information we are able to obtain knowledge about the *indentation area* (m-) and *protrusion area* (M+) of an object, as defined in the *Process-Grammar* [25]. The details of the algorithm produced for the determination of the object's convexity-type can be found in Appendix A.1.

Unsupervised Learning from *AsGs*. Clustering AsGs results in groups of similar pair-wise interaction graphs. Having extracted a set of AsGs, we measure the difference between these graph structures, and perform divisive clustering [26] on them to obtain classes of interactions, and thus classes of affordances. From the produced classes of interactions, obtained from processing only object-object interactions, we can infer the presence of affordances which are similar to each other.

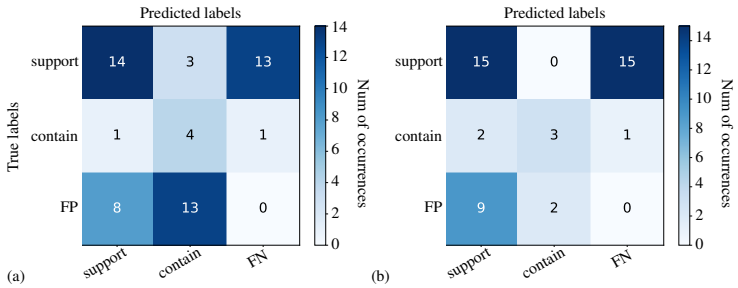
To measure the difference between AsGs we consider the differences of their spatial and temporal vertex sets (V^s_{spat}, V^s_{temp}). $V^s_{entities}$ is not exploited in this process since each object is represented by a class-agnostic entity, thus does not contribute any information. Let G^α and G^β be two AsGs and $V^{s\alpha}_{spat}$, $V^{s\alpha}_{temp}$, $V^{s\beta}_{spat}$, and $V^{s\beta}_{temp}$ the spatial and temporal vertex sets respectively. The set of the unique spatial ($V^{s\alpha,s\beta}_{spat}$) and temporal ($V^{s\alpha,s\beta}_{temp}$) vertices comprise of the vertices of the two graphs which are present in only one of the G^α and G^β, as defined in Eq. 1.

$$V^{s\alpha,s\beta}_R = \{v : v \in \{V^{s\alpha}_R \setminus V^{s\beta}_R\} \cup \{V^{s\beta}_R \setminus V^{s\alpha}_R\}\}, \text{where } R \in \{spat, temp\} \tag{1}$$

To cluster relational graph structures we employ a metric of vertex differences for capturing similarities and dissimilarities between graphs. As the relational graph structures we cluster are undirected and consist of specific types of vertices, spatial and temporal, we use a *set Edit Distance* (sED) metric with equally weighted intra-vertex-type differences, which is defined as:

$$sED = \frac{1}{2}\sum\nolimits_{v \in V^{s\alpha,s\beta}_{spat}} v + \frac{1}{2}\sum\nolimits_{v \in V^{s\alpha,s\beta}_{temp}} v \tag{2}$$

where $V^{s\alpha,s\beta}_{spat}$ and $V^{s\alpha,s\beta}_{temp}$ are defined in Eq. 1. Graphs which represent different interactions, hence their spatial and temporal relational sets differ, have a higher value of the sED metric from graphs representing the same or similar interaction. We demonstrate that by utilizing the introduced D-RCC5 relations more complete and homogeneous clusters are formed enhancing the robustness of the overall affordance clustering system.



Metrics	Baseline	Ours
V-measure	0.2686	**0.3433**
NMI	0.2838	**0.3539**
Homogeneity	0.3968	**0.4532**
Completeness	0.2030	**0.2763**

(c)

Figure 3: Enhanced confusion matrices without (a) and with (b) using D-RCC5 relations, as well as metrics of the baseline and our approach (c).

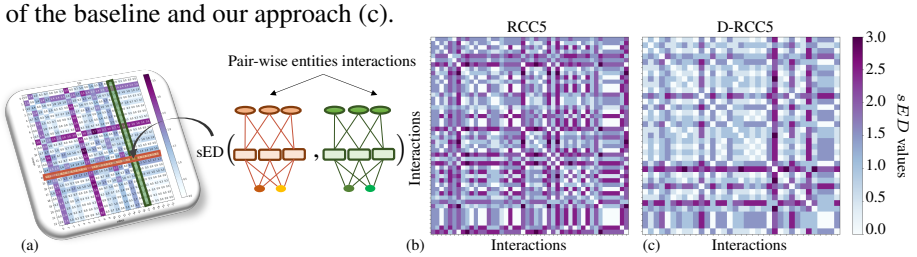


Figure 4: sED values of all possible comparisons of pair-wise interactions.

3 Experiments

We evaluate the proposed approach on the CAD120 dataset [27]. Class-agnostic objects' bounding boxes are acquired by employing the Faster R-CNN framework [28] with a pre-trained model on COCO dataset [29] using the ResNet101 network architecture [30]. We further exploited the Convolutional Pose Machine [31] to obtain the skeletal data of the humans in the scene. Additionally, to eliminate the human's RGB-D information from the predicted objects' bounding boxes, we exploit a human mask from the DensePose R-CNN framework [32]. Further details of the experimental setup can be found in Appendix A.2. We compare our work with a baseline, which comprises relational graphs with standard RCC5 spatial relations [22, 23], similar to [19].

For the evaluation of the clusters produced, we exploit the *v-measure*, *homogeneity*, *completeness*, and *normalized-mutual information* scores. Also we analyze the results of an *enhanced confusion matrix*, which is defined later in this section. All metrics are normalized, with higher scores corresponding to better correlation of the predicted clusters to the ground truth classes of affordances. Fig. 3(c) summarizes our results, indicating an increase in all examined metrics demonstrating notable benefits obtained by incorporating objects' depth information in the AGs.

Furthermore, we compare enhanced confusion matrices for the baseline and the proposed approach, featuring an extra row and column for the false positive (FP) and false negative (FN) interactions of objects accordingly. Specifically, for every predicted affordance cluster we observe the value of interactions that do not correspond to a true affordance label (FP); we also examine for every affordance class the value of non-detected interactions (FN). From the enhanced confusion matrix we obtain a 47.6% decrease of the FP object interactions while employing D-RCC5 relations. In Fig. 3 a clear distinction is illustrated between the affordance classes when D-RCC5 spatial relations are being employed, in contrast to the baseline where the 'support' affordance can be misclassified as 'contain'. However, Fig. 3(a) introduces a notable score of misclassification of the 'contain' affordance to the class 'support'. By analyzing the test set employed to evaluate our approach we observe that a significant number of concave objects were not detected as the concavity was not visible due to the camera's position. This resulted to a 16.6 percentage points drop of classification accuracy for the 'contain' affordance causing a 33% of 'contain' affordances to be classified as 'support'.

Fig. 4 illustrates the graph differences between all possible pair-wise interactions detected. Darker cell colors correspond to a higher sED value, indicating more homogeneous clusters with D-RCC5 relations. More qualitative results can be found in Appendix 6.

4 Conclusions

We have addressed the problem of learning in an unsupervised way graph structures from RGB-D video data to predict domain independent object affordances. Our experiments demonstrate that enhancing the Activity Graphs with the objects' depth information produces more complete clusters of affordances compared to primitive spatio-temporal relations.

References

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

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

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

[4] Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly Supervised Affordance Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2795–2804, 2017.

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

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

[7] Luis Montesano and Manuel Lopes. Learning Grasping Affordances from Local Visual Descriptors. In *2009 IEEE 8th international conference on development and learning*, pages 1–6. IEEE, 2009.

[8] Yibiao Zhao and Song-Chun Zhu. Scene Parsing by Integrating Function, Geometry and Appearance Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3119–3126, 2013.

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

[10] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and Recognizing Human-Object Interactions. *arXiv preprint arXiv:1704.07333*, 2017.

[11] Bangpeng Yao, Jiayuan Ma, and Li Fei-Fei. Discovering Object Functionality. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2512–2519, 2013.

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

[13] Yu Sun, Shaogang Ren, and Yun Lin. Object–object interaction affordance learning. *Robotics and Autonomous Systems*, 62(4):487–496, 2014.

[14] Shaogang Ren and Yu Sun. Human-Object-Object-Interaction Affordance. In *Robot Vision (WORV), 2013 IEEE Workshop on*, pages 1–6. IEEE, 2013.

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

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

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

[18] Shane Griffith, Vladimir Sukhoy, and Alexander Stoytchev. Using Sequences of Movement Dependency Graphs to Form Object Categories. In *Humanoid Robots (Humanoids), 2011 11th IEEE-RAS International Conference on*, pages 715–720. IEEE, 2011.

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

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

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

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

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

[24] James F Allen. Maintaining Knowledge about Temporal Intervals. In *Readings in qualitative reasoning about physical systems*, pages 361–372. Elsevier, 1990.

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

[26] Leonard Kaufman and Peter J Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*, volume 344. John Wiley & Sons, 2009.

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

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

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

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[31] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional Pose Machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

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

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

A Appendices

A.1 Spatial Relation Depth-enhancement Algorithm

Alg. 1 describes the convexity-type selection process for each object detected in the scene. The algorithm uses as input the depth distribution from the predicted object's bounding box and outputs its convexity-type. The human's RGB-D information is excluded from the processed bounding box, employing a deep learning-based human pose estimation framework, for acquiring more accurate concave curve information. Initially a convexity threshold has been selected ($thresh_{convex}$) defining the upper limit of the depth range of a 'convex'-type object, the selection of which is discussed in Appendix A.2. Objects with depth range exceeding this threshold are either considered 'concave'- or 'surface'-type depending on their depth contour hierarchies. Contour hierarchies provide a tree structure of contour inclusion, where every node of the tree stands for a contour and every parent includes its children. Hence, the detection of a child contour in the depth space, deduces the presence of a concave curve , thus a 'concave'-type object and 'surface'-type otherwise.

The determination of the object's convexity-type serves in defining the boundaries of the object's depth information, as presented in Alg. 2. The algorithm exploits the predicted object's depth distribution to infer its depth boundaries. The object's depth is employed to ascertain an RCC5 relation when two objects interact in the 2D image plane, *e.g.* the PP RCC5 spatial relation occurs between one or more 'concave'-type objects when the depth information of the containee object confirms that it is between the m- and M+ areas of the container object. For a 'concave'-type object we consider partitioning the depth information into h sections where the n with the highest depth values are estimated to capture the concave curve of the object. We set the depth boundaries of such objects to enclose the concave curve's depth information for detecting a PP RCC5 relation. For a 'convex'- and 'surface'-type object the depth boundaries are not processed since no concavity is present.

Algorithm 1 Define convexity type of object.

Given: $thresh_{convex}$

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

Algorithm 2 Define convexity area of an object.

Given: h, n

procedure CONVEXITYAREA($\text{dist}_{\text{depth}}$)
  $max_d \leftarrow max(dist_{depth})$
  $min_d \leftarrow min(dist_{depth})$
  $object_{type} \leftarrow$ ObjectConvexity($dist_{depth}$)
  if $object_{type} =$ concave **then**
    $sections \leftarrow (max_d - min_d)/h$
    $max_{cArea} \leftarrow max_d$
    $min_{cArea} \leftarrow max_d - (n * sections)$
  else
    $max_{cArea} \leftarrow max_d$
    $min_{cArea} \leftarrow min_d$
  return max_{cArea}, min_{cArea}
  end

A.2 Experimental Setup

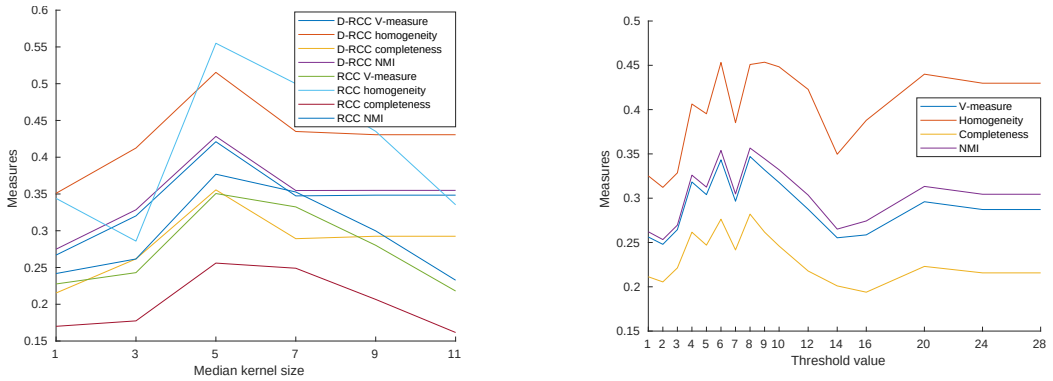
The methodology introduced was evaluated on the CAD120 dataset [27] which consists of 120 videos, of 30fps, capturing human activities from everyday-life scenarios. Each video records a single activity with a single actor, who interacts with one or more objects relevant to the activity being performed. For the experiments conducted a test set of 10% of the total amount of videos was randomly selected, while the rest was used as training set.

Class-agnostic objects' bounding boxes were acquired by employing the Faster R-CNN framework [28]. The set of object classes used for detection consist of all indoor manipulable objects from the COCO dataset. To filter out false positive and multiple detections of objects, we set the detection confidence score to 50% and the value of the Non-maximum suppression metric to 0.2, selected from an empirical study. Furthermore, we employed the QSRlib library [33] for the construction of the AGs.

However, even when employing qualitative spatial relations to represent spatial interactions of objects, the transitions from one relation to another is not always smooth. To eliminate sparse and false

positive spatial relations we exploit a median filter of kernel size λ acting on the qualitative spatial relation detection sequence. Fig. 5(a) illustrates the experiments conducted on the training set with various kernel sizes ($\lambda \in \{1, 3, 5, 7, 9, 11\}$) for determining the best value for the kernel size in reference to the reported evaluation metrics. A median filter with kernel size 5 was applied.

Fig. 5(b) demonstrates the experiments on the training set for the selection of the $thresh_{convex}$ threshold value employed for conditioning the 'convex'-type objects, as described in Appendix A.1. Regarding the examined metrics, a value of $thresh_{convex} = 6$ achieves higher scores in all measures.



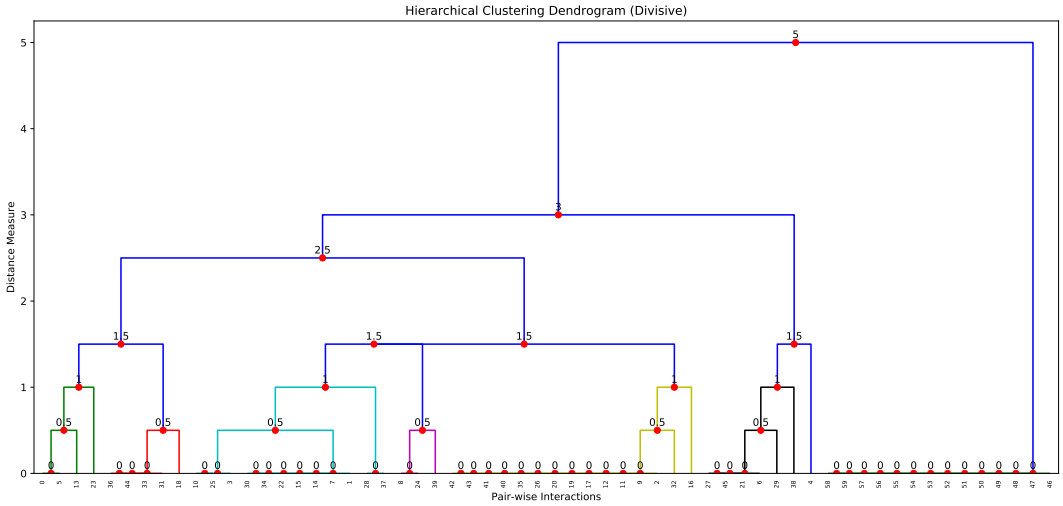
(a) (b)

Figure 5: Tuning of the kernel size (λ) of the median filter and $thresh_{convex}$ for defining 'convex'-type objects, respectively.

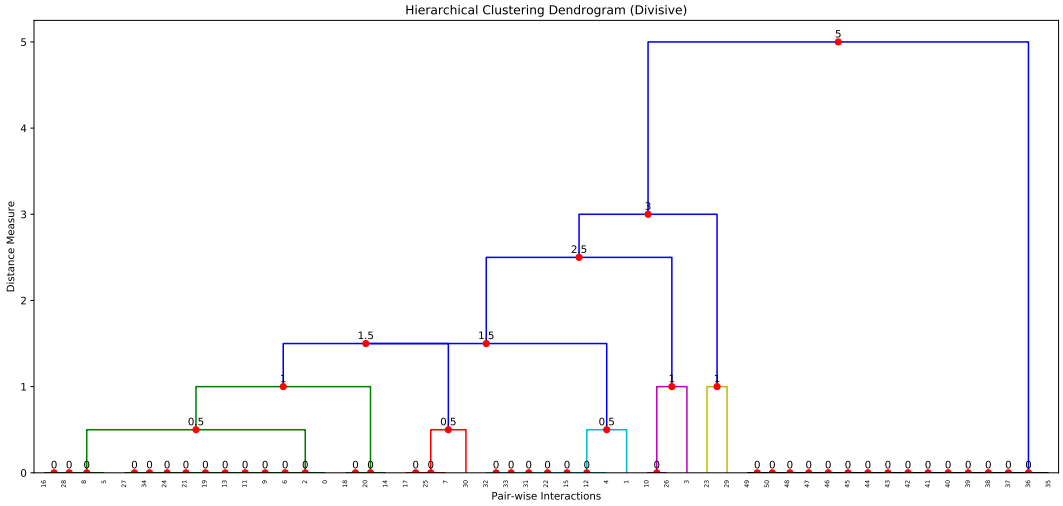
A.3 Qualitative Results

The points in the x-axis of the depicted dendrograms in Fig. 6 correspond to the pair-wise interactions detected and the y-axis to the distance measure used for clustering these interactions. The colored branches of the dendrograms demonstrate the different interaction classes corresponding to distinct clusters. Though we evaluate two prominent affordances ('contain' and 'support'), multiple clusters are being produced from the clustering mechanism. This results from differences in the AGs of the same described affordances because they were performed in different ways by different agents.

Fig. 6 demonstrates that many small clusters, produced when employing RCC5 relations, are merged into a single one, when exploiting D-RCC5. Consequently, the overall number of clusters is reduced leading to a higher completeness clustering result.



(a) With RRC5 relations.



(b) With D-RCC5 relations.

Figure 6: Sample outputs of the employed divisive clustering mechanism.