





One-Shot Learning for Human Affordance Detection

Abel Pacheco-Ortega¹  and Walterio Mayol-Cuevas^{1,2} 

¹ University of Bristol, Bristol, UK

{abel.pachecoortega,walterio.mayol-cuevas}@bristol.ac.uk

² Amazon, Seattle, WA 98109, USA

Abstract. The diversity of action possibilities offered by an environment, a.k.a affordances, cannot be addressed in a scalable manner simply from object categories or semantics, which are limitless. To this end, we present a one-shot learning approach that trains on one or a handful of human-scene interaction samples. Then, given a previously unseen scene, we can predict human affordances and generate the associated articulated 3D bodies. Our experiments show that our approach generates physically plausible interactions that are perceived as more natural in 60–70% of the comparisons with other methods.

Keywords: Scene understanding · Affordances detection · Human interactions · Visual perception · Affordances

1 Introduction

Coined by James J. Gibson in [3], affordances refer to the action possibilities offered by the environment to an agent. He claimed that living beings perceive their environment in terms of such affordances.

An artificial agent with object, semantics and human affordances detection capabilities would be able to identify elements, their relations and the locations in the environment that support the execution of actions like stand-able, walk-able, place-able, and sit-able. This enhanced scene understanding is helpful in the Metaverse, where virtual agents should execute actions or where scenes must be populated by humans performing a given set of interactions.

We present a direct representation of human affordances that extracts a meaningful geometrical description through analysing proximity zones and clearance space between interacting entities in human-environment configurations.



Fig. 1. Trained in a one-shot manner, our approach detects human affordances and hallucinates the associated human bodies interacting with the environment in a natural and physically plausible way

Our approach can determine locations in the environment that support them and generate natural and physically plausible 3D representations (see Fig. 1). We compare our method with state-of-the-art intensively trained methods.

2 Related Work

Popular interpretations of the concept of affordances refer to them as *action possibilities* or *opportunities of interaction* for an agent/animal that are perceived directly from the shape and form of the environment/object.

The affordances detection from RGB images was explored by Gupta et al. [4] with a voxelised geometric estimator. Lately, data-intensive approaches were used by Fouhey et al. [2] with a detector trained with labels on RGB frames from the NYUv2 dataset [13] and by Luddecke et al. [9] with a residual neural network trained with a lookup table between affordances and objects parts on the ADE20K dataset [20].

Other approaches go further by synthesising the detected human-environment interaction. The representation of such interactions has been showcased with human skeletons in [7, 8, 15]; nevertheless, their representativeness cannot be reliably evaluated because contacts, collisions, and the naturalness of human poses are not entirely characterised.

Closer to us, efforts with a more complex interaction representation over 3D scenes have been explored. Ruiz and Mayol [12] developed a geometric interaction descriptor for non-articulated, rigid object shapes with good generalisation in detecting physically feasible interaction configurations. Using the SMPL-X human body model [10], Zhang et al. [18] developed a context-aware human body generator that learnt the distribution of 3D human poses conditioned on the depth and semantics of the scene from recordings in the PROX dataset [5]. In a follow-up effort, Zhang et al. [17] developed a purely geometric approach to model human-scene interactions by explicitly encoding the proximity between the body and the environment, thus only requiring a mesh as input. Lately, Hassan et al. in [6] learnt the distribution of contact zones in human body poses and used them to find environment locations that better support them.

Our main difference from [5, 6, 17, 18] is that ours is not a data-driven approach; ours does not require the use of most, if not all, of a labelled dataset, e.g. around 100K image frames in PROX [5]. Just one if not a few examples of interactions are necessary to train our detector, as in [12], but we extend the descriptor to consider the clearance space of the interactions and their uses and optimise with the SMPL-X human model after positive detection.

3 Method

3.1 A Spatial Descriptor for Spatial Interactions

Inspired by recently developed methods that have revisited geometric features such as the bisector surface for scene-object indexing [19] and affordance detection [12], our affordance descriptor (see Fig. 2) expands on the Interaction Bisector Surface (IBS) [19], an approximation of the well-known Bisector Surface (BS)

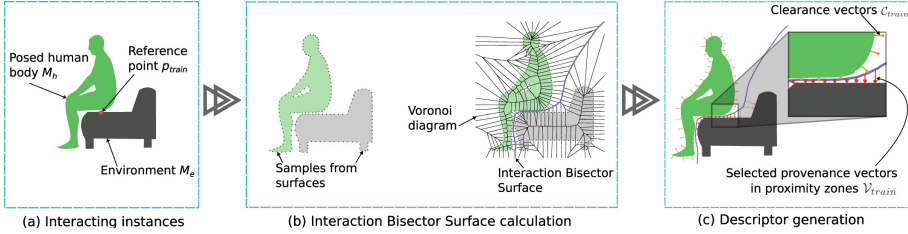


Fig. 2. Illustrative 2D representation of our training pipeline. (a) Given a posed human body M_h interacting with an environment M_e on a reference point p_{train} , (b) we extract the Interaction Bisector Surface from the Voronoi diagram of sample points on M_h and M_e , then (c) we use the IBS to characterise the proximity zones and the surrounding space with provenance and clearance vectors

[11]. Given two surfaces $S_1, S_2 \in \mathbb{R}^3$, the BS is the set of sphere centres that touch both surfaces at one point each.

Our one-shot training process requires 3-tuples (M_h, M_e, p_{train}) , where M_h is a posed human body mesh, M_e is an environment mesh and p_{train} is a reference point on M_e where the interaction is supported.

Let P_h and P_e denote the sampling points on M_h and M_e , respectively; their IBS \mathcal{I} is defined as:

$$\mathcal{I} = \left\{ p \mid \min_{p'_h \in P_h} \|p - p'_h\| = \min_{p'_e \in P_e} \|p - p'_e\| \right\} \tag{1}$$

We operate the Voronoi diagram \mathcal{D} generated with P_h and P_e to produce \mathcal{I} . By construction, every ridge in \mathcal{D} is equidistant to the couple of points that define it. Then, \mathcal{I} is composed of ridges in \mathcal{D} generated because of points from both P_h and P_e . An IBS can reach infinity, but we limit \mathcal{I} by clipping it with the bounding sphere of M_h augmented $ibs_{r,f}$ times in its radius. A low sampling rate degenerates on an IBS that pierces the boundaries of M_h or M_e . A higher density of samples is critical in those zones where the proximity between the interacting meshes is small. We use three stages to populate P_h and P_e : 1) We generate Poisson disk sample sets [16] of ibs_{ini} points on each M_e and M_h . 2) *Counterpart sampling* strategy. We append to P_e the closest points on M_e to elements in P_h , and equally, we integrate into P_h the closest point on M_h to samples in P_e . We executed the *counterpart sampling* strategy ibs_{cs} times. 3) *Collision point sampling* strategy. We calculate a preliminary IBS and test it for collisions with M_h and M_e ; if they exist, we add as samples the points where collisions occur as well as their counterpart points. We perform the *collision point sampling* strategy until we get an IBS that does not pierce M_h nor M_e .

To capture the regions of interaction proximity on our enhanced IBS, as mentioned above, we use the notion of provenance vectors [12]. The *provenance vectors* of an interaction start from any point on \mathcal{I} and finish at the nearest point on M_e . Formally:

$$V_p = \{(a, \vec{v}) \mid a \in \mathcal{I}, \vec{v} = \arg \min_{e \in M_e} \|e - a\| - a\} \quad (2)$$

where a is the starting point of the delta vector \vec{v} to the nearest point on M_e . *Provenance vectors* inform about the direction and distance of the interaction; the smaller the vector, the more noteworthy is for the description of the interaction. Let $V'_p \subset V_p$ the subset of *provenance vectors* that ends at any point in P_e ; we perform a weighted randomised sampling selection of elements from V'_p with the weight allocation as follows:

$$w_i = 1 - \frac{|\vec{v}_i| - |\vec{v}_{min}|}{|\vec{v}_{max}| - |\vec{v}_{min}|}, \quad i = 1, 2, \dots, |P_e| \quad (3)$$

where $|\vec{v}_{max}|$ and $|\vec{v}_{min}|$ are the norms of the biggest and smallest vectors in V'_p respectively. The selected *provenance vectors* \mathcal{V}_{train} integrate into our affordance descriptor with an adjustment to normalise their positions with the defined reference point p_{train} :

$$\mathcal{V}_{train} = \{(a'_i, \vec{v}_i) \mid a'_i = a_i - p_{train}, i = 1, 2, \dots, num_{pv}\} \quad (4)$$

where num_{pv} is the number of samples from V'_p to integrate into our descriptor.

However, the *provenance vectors* are, on their own, insufficient to capture the whole nature of the interaction on highly articulated objects such as the human body. We expand this concept by taking a more comprehensive description that includes a set of vectors to define the surrounded space necessary for the interaction. Given S_H an evenly sampled set of num_{cv} points on M_h , the *clearance vectors* that integrate to our descriptor \mathcal{C}_{train} are defined as follows:

$$\mathcal{C}_{train} = \{(s'_j, \vec{c}_j) \mid s'_j = s_j - p_{train}, s_j \in S_H, \vec{c}_j = \psi(s_j, \hat{n}_j, \mathcal{I})\} \quad (5)$$

$$\psi(s'_j, \hat{n}_j, \mathcal{I}) = \begin{cases} d_{max} \cdot \hat{n}_j & \text{if } \varphi(s_j, \hat{n}_j, \mathcal{I}) > d_{max} \\ \varphi(s_j, \hat{n}_j, \mathcal{I}) \cdot \hat{n}_j & \text{otherwise} \end{cases} \quad (6)$$

where p_{train} is the defined reference point, \hat{n}_i is the unit surface normal vector on sample s_j , d_{max} is the maximum norm of any \vec{c}_j , and $\varphi(s_j, \hat{n}_j, \mathcal{I})$ is the distance travelled by a ray with origin s_j and direction \hat{n}_i until collision with \mathcal{I} .

Formally, our affordances descriptor is defined as:

$$f : (M_h, M_e, p_{train}) \longrightarrow (\mathcal{V}_{train}, \mathcal{C}_{train}, \hat{n}_{train}) \quad (7)$$

where \hat{n}_{train} is the unit surface normal vector of M_e at p_{train} .

3.2 Human Affordances Detection

Let $\mathcal{A} = (\mathcal{V}_{train}, \mathcal{C}_{train}, \hat{n}_{train})$ an affordance descriptor, we define its rigid transformations as:

$$\Omega(\mathcal{A}, \phi, \tau) = (\mathcal{V}_{\phi\tau}^A, \mathcal{C}_{\phi\tau}^A, \hat{n}_{train}) \quad (8)$$

$$\mathcal{V}_{\phi\tau}^A = \{(a''_i, \vec{v}_i) \mid a''_i = R_\phi \cdot a'_i + \tau, (a'_i, \vec{v}_i) \in \mathcal{V}_{train}\}$$

$$\mathcal{C}_{\phi\tau}^A = \{(s''_i, \vec{c}_i) \mid b''_i = R_\phi \cdot s'_i + \tau, (s'_i, \vec{c}_i) \in \mathcal{C}_{train}\}$$

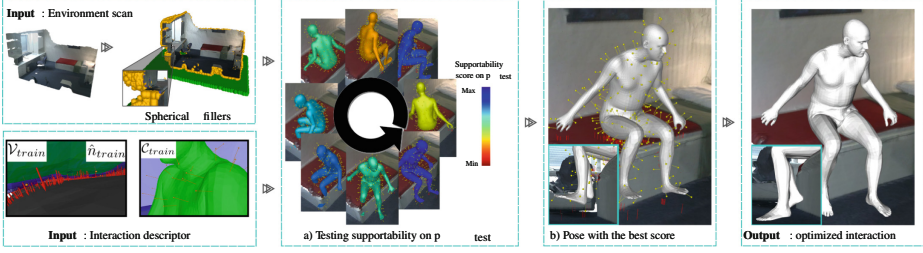


Fig. 3. We determine supportability of interaction on a given point by (a) measuring compatibility of surface normal, as well as provenance and clearance vector over different rotated configurations. (b) After a positive detection, the body pose is optimised to generate a natural and physically plausible interaction

where $\tau \in \mathbb{R}^3$ is the translation vector, ϕ is the rotation around z defined by the rotation matrix R_ϕ .

We determine that a test location p_{test} on an environment M_{test} with a unit surface normal vector \hat{n}_{test} supports a trained interaction \mathcal{A} if the angle difference between \hat{n}_{train} and \hat{n}_{test} is less than a threshold $\rho_{\hat{n}}$, and its translated descriptor at p_{test} has a good alignment of provenance vectors and a gated number of clearance vector that collide with M_{test} in any of the n_ϕ different ϕ values used during the test.

After corroborating the match between train and test normal vectors, we transform the interaction descriptor \mathcal{A} with $\tau = p_{test}$ and n_ϕ different ϕ values within $[0, 2\pi]$. For each calculated 3-tuple $(\mathcal{V}_{\phi\tau}^A, \mathcal{C}_{\phi\tau}^A, \hat{n}_{train})$, we generate a set of rays R_{pv} defined as follows:

$$R_{pv} = \left\{ (a''_i, \hat{v}_i) \mid \hat{v}_i = \frac{\vec{v}_i}{\|\vec{v}_i\|}, (a''_i, \vec{v}_i) \in \mathcal{V}_{\phi\tau}^A \right\} \quad (9)$$

where a''_i is the starting point, and $\hat{v}_i \in \mathbb{R}^3$ is the direction of each ray. Then we extend each ray in R_{pv} by ϵ_i^{pv} until collision with M_{test} as

$$(a'' + \epsilon_i^{pv} \cdot \hat{v}_i) \in M_{test}, \quad i = 1, 2, \dots, num_{pv} \quad (10)$$

and compare with the magnitude of each correspondent provenance vector in $\mathcal{V}_{\phi\tau}^A$. When any element in R_{pv} extends beyond a predetermined limit max_{long} , the collision with the environment is classified as non-colliding. We calculate the alignment score κ as a sum of the differences between the extended rays and the trained *provenance vectors* with

$$\kappa = \sum_{\forall i | \epsilon_i^{pv} \leq max_{long}} |\epsilon_i^{pv} - \vec{v}_i| \quad (11)$$

The higher the κ value, the less supportability of the interaction on p_{test} . We experimentally determine interaction-wise thresholds for the sum of differences



Fig. 4. Action planning as a further step. Left: 3 affordances evaluated in an environment. Right: scores are used to plan concatenated action milestones

max_{κ} and the number of missing ray collisions $max_{missings}$ that allow us to score the affordance capabilities on p_{test} .

Clearance vectors are meant to fast detect collision configurations by the calculation of ray-mesh intersections. Similarly to *provenance vectors*, we generate a set of rays R_{cv} with origins and directions determined by $\mathcal{C}_{\phi\tau}^A$. We extend the rays in R_{cv} until collision with the environment and calculate its extension ϵ_j^{cv} . Extended rays with $\epsilon_j^{cv} \leq \|\vec{c}_j\|$ are considered as possible collisions. In practise, we also track an interaction-wise threshold to refuse supportability due to collisions $max_{collisions}$. A sparse distribution of clearance vectors on noisy meshes results in collisions not detected by *clearance vectors*. To improve, we enhance scenes with a set of *spherical fillers* that pad the scene (see Fig. 3).

Every human-environment interaction trained from the PROX dataset [5] has an associated SMPL-X characterisation that we use to optimise the human pose with previously determined body contact regions, the AdvOptim loss function presented in [17] and the SDF values of the scene.

4 Experiments

We evaluate the physical plausibility and the perception of the naturalness of the human-environment interactions generated. Our baselines are the approaches presented in PLACE [17] and POSA [6].

PROX [5] is a dataset with 12 scanned indoor environments and 20 recordings with data of subjects interacting within them. We divide PROX into train and test sets following the setup in [17]. To generate our descriptors, we get data from 23 manually selected frames with subjects sitting, standing, reaching, lying, and walking. We also test on 7 rooms from MP3D [1] and 5 rooms of Replica [14].

We generate the IBS surface \mathcal{I} with an initial sampling set of $ibs_{ini} = 400$ points on each surface, with the *counterpart sampling* strategy executed $ibs_{cs} = 4$ times and a cropping factor of $ibs_{rf} = 1.2$. Our descriptors are made up of $num_{pv} = 512$ *provenance vectors* and $num_{cv} = 256$ *clearance vectors* extended up to $d_{max} = 5[cm]$. In testing, we use a normals angle difference threshold of $\rho_{\vec{n}} = \pi/3$, check for supportability on $n_{\phi} = 8$ different directions and extend *provenance vectors* up to $max_{long} = 1.2$ times the sphere radius used for cropping \mathcal{I} during training.

Physical Plausibility Test. We use the non-collision and contact scores as in [17], but include an additional cost metric that indicates the collision depth

Table 1. Physical plausibility. Non collision, contact and collision depth scores (\uparrow : benefit, \downarrow : cost) before and after optimization. Best results boldface

Model	Optimizer	Non collision \uparrow			Contact \uparrow			Collision depth \downarrow		
		PROX	MP3D	Replica	PROX	MP3D	Replica	PROX	MP3D	Replica
PLACE	w/o	0.9207	0.9625	0.9554	0.9125	0.5116	0.8115	1.6285	0.8958	1.2031
PLACE	SimOptim	0.9253	0.9628	0.9562	0.9263	0.5910	0.8571	1.8169	1.0960	1.5485
PLACE	AdvOptim	0.9665	0.9798	0.9659	0.9725	0.5810	0.9931	1.6327	1.1346	1.6145
POSA (contact)	w/o	0.9820	0.9792	0.9814	0.9396	0.9526	0.9888	1.1252	1.5416	2.0620
POSA (contact)	optimized	0.9753	0.9725	0.9765	0.9927	0.9988	0.9963	1.5343	2.0063	2.4518
Ours	w/o	0.9615	0.9853	0.9931	0.5654	0.3287	0.4860	0.1648	0.1326	0.2096
Ours	AdvOptim	0.9816	0.9853	0.9883	0.9363	0.6213	0.8682	0.6330	0.8716	0.8615

between the generated body and the scene. We generate 1300 interacting bodies per model in each scene and report the averages of the scores in Table 1. In all datasets, bodies generated with our optimised model present high non-collision as well as low contact and collision-depth scores.

Perception of Naturalness Test. Every scene in our datasets is used equally in the random selection of 162 test locations. We use the optimised version of the models to generate human-environment interactions at test locations and evaluate their perceived naturalness on Amazon Mechanical Turk. Each MTurk performs 11 randomly selected assessments, including two control questions, by observing interactions with dynamic views. Three different MTurks evaluate every item. In a side-by-side evaluation, we simultaneously present outputs from two different models. Answers to “Which example is more natural?” show that our human-environment configurations are preferred on 60.7% and 72.6% of the comparisons with PLACE and POSA, respectively. In an individual evaluation, where every interaction generated is assessed with the question “The human is interacting very naturally with the scene. What is your opinion?” using a 5-point Likert scale (from 1 for “strongly disagree” to 5 for “strongly agree”), the mean and standard deviations of the evaluations are: PLACE 3.23 ± 1.35 , POSA 2.79 ± 1.18 , and ours 3.39 ± 1.25 .

5 Conclusion

Our approach generalises well to detect interactions and generate natural and physically plausible body-scene configurations. Understanding a scene in terms of action possibilities is a desirable capability for autonomous agents performing in the Metaverse (see Fig. 4).

Acknowledgments. Abel Pacheco-Ortega thanks the Mexican Council for Science and Technology (CONACYT) for the scholarship provided for his studies with the scholarship number 709908. Walterio Mayol-Cuevas thanks the visual egocentric research activity partially funded by UK EPSRC EP/N013964/1.

References

1. Chang, A., et al.: Matterport3D: learning from RGB-D data in indoor environments. In: International Conference on 3D Vision (3DV) (2017)
2. Fouhey, D.F., Wang, X., Gupta, A.: In defense of the direct perception of affordances. arXiv preprint [arXiv:1505.01085](https://arxiv.org/abs/1505.01085) (2015). <https://doi.org/10.1002/eji.201445290>
3. Gibson, J.J.: The theory of affordances. In: *Perceiving, Acting and Knowing. Toward and Ecological Psychology*. Lawrence Erlbaum Associates (1977)
4. Gupta, A., Satkin, S., Efros, A.A., Hebert, M.: From 3D scene geometry to human workspace. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1961–1968, June 2011. IEEE. <https://doi.org/10.1109/CVPR.2011.5995448>. <http://ieeexplore.ieee.org/document/5995448/>
5. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constraints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2282–2292 (2019)
6. Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3D scenes by learning human-scene interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14708–14718 (2021)
7. Jiang, Y., Koppula, H.S., Saxena, A.: Modeling 3D environments through hidden human context. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**, 2040–2053 (2016). <https://doi.org/10.1109/TPAMI.2015.2501811>
8. Li, X., Liu, S., Kim, K., Wang, X., Yang, M.H., Kautz, J.: Putting humans in a scene: learning affordance in 3D indoor environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12368–12376 (2019)
9. Luddecke, T., Worgotter, F.: Learning to segment affordances. In: The IEEE International Conference on Computer Vision (ICCV) Workshops, pp. 769–776. IEEE, October 2017. <https://doi.org/10.1109/ICCVW.2017.96>. <http://ieeexplore.ieee.org/document/8265305/>
10. Pavlakos, G., et al.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10975–10985 (2019)
11. Peternell, M.: Geometric properties of bisector surfaces. *Graph. Models* **62**(3), 202–236 (2000). <https://doi.org/10.1006/gmod.1999.0521>
12. Ruiz, E., Mayol-Cuevas, W.: Geometric affordance perception: leveraging deep 3D saliency with the interaction tensor. *Front. Neurobot.* **14**, 45 (2020)
13. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) *ECCV 2012*. LNCS, vol. 7576, pp. 746–760. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_54
14. Straub, J., et al.: The replica dataset: a digital replica of indoor spaces. arXiv preprint [arXiv:1906.05797](https://arxiv.org/abs/1906.05797) (2019)
15. Wang, X., Girdhar, R., Gupta, A.: Binge watching: scaling affordance learning from sitcoms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2596–2605 (2017)
16. Yuksel, C.: Sample elimination for generating poisson disk sample sets. In: *Computer Graphics Forum*, vol. 34, pp. 25–32 (2015). <https://doi.org/10.1111/cgf.12538>
17. Zhang, S., Zhang, Y., Ma, Q., Black, M.J., Tang, S.: PLACE: proximity learning of articulation and contact in 3D environments. In: 8th International Conference on 3D Vision (3DV 2020) (virtual) (2020)

18. Zhang, Y., Hassan, M., Neumann, H., Black, M.J., Tang, S.: Generating 3D people in scenes without people. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020). <https://github.com/yz-cnsdq/PSI-release/>
19. Zhao, X., Wang, H., Komura, T.: Indexing 3D scenes using the interaction bisector surface. *ACM Trans. Graph.* **33**(3), 1–14 (2014). <https://doi.org/10.1145/2574860>. <http://dl.acm.org/citation.cfm?doid=2631978.2574860>
20. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)