What Makes a Chair a Chair?

Helmut Grabner[1] Juergen Gall[1] Luc Van Gool[1,2]

[1]Computer Vision Laboratory
ETH Zurich
{grabner,gall,vangool}@vision.ee.ethz.ch

[2]ESAT - PSI / IBBT
K.U. Leuven
luc.vangool@esat.kuleuven.be

Abstract

Many object classes are primarily defined by their functions. However, this fact has been left largely unexploited by visual object categorization or detection systems. We propose a method to learn an affordance detector. *It identifies locations in the 3d space which "support" the particular function. Our novel approach "imagines" an actor performing an action typical for the target object class, instead of relying purely on the visual object appearance. So, function is handled as a cue complementary to appearance, rather than being a consideration after appearance-based detection. Experimental results are given for the functional category "sitting". Such affordance is tested on a 3d representation of the scene, as can be realistically obtained through SfM or depth cameras. In contrast to appearance-based object detectors, affordance detection requires only very few training examples and generalizes very well to other sittable objects like benches or sofas when trained on a few chairs.*

1. Introduction

"An object is first identified as having important functional relations, [...] perceptual analysis is derived of the functional concept [...]."

Nelson, 1974, [17]

"Affordances relate the utility of things, events, and places to the needs of animals and their actions in fulfilling them [...]. Affordances themselves are perceived and, in fact, are the essence of what we perceive."

Gibson, 1982, [8, p. 60]

"There's little we can find in common to all chairs – except for their intended use."

Minsky, 1986, [16, p. 123]

"[...] objects like coffee cups are artifacts that were created to fulfill a function. The function of an object plays a critical role in processing that object [... for] categorization and naming."

Carlson-Radvansky *et al.*, 1999, [4]

Figure 1. The "chair-challenge" by I. and H. Bülthoff [3] (reprint with the author's permission).

These quotes emphasize that functional properties or affordances[1] are essential for forming concepts and learning object categories. Experiments (*e.g.* [18, 4]) have demonstrated that both appearance and function are strong cues for learning by infants. Initially they attend only to the form of an object. Later they use form and function and finally (by the age of 18 months) they attend to the relationships between form and function. Furthermore, Booth and Waxman [2] have identified two salient cues that facilitate categorization in infancy, namely (i) object functions and (ii) object names. Moreover, names of objects most often evolve on the basis of function[2].

Whereas all this is well known for a long time, it has been left mostly unused for object detection in computer vision. Taking a look at the results of the recent Pascal VOC Challenge [5], the performance still strongly depends

[1]"*Affordance:* A situation where an object's sensory characteristics intuitively imply its functionality and use. [...] A chair, by its size, its curvature, its balance, and its position, suggests sitting on it.", http://www.usabilityfirst.com/glossary/affordance, 2010/07/28. Introduced in 1979 by Gibson [9, p. 127] based on the verb *afford*.

[2]When considering the evolution of a word for an object, most of the time it is based on its function. For example the word "chair": PIE base **sed-* (to sit) → Latin *sedentarius* (sitting, remaining in one place) → *sedentary* (meaning "not in the habit of exercise") → *cathedral* → *chair*. http://www.etymonline.com, 2010/10/02.

Figure 2. Visual object categories vs. functional categories. In this work, we are interested in observing the functional category, hence our detector is rather a functionality detector than a traditional object detector.

on the object class. Whereas categories like "airplanes" and "bicycles" are handled with reasonable success, other categories like "chairs" fare less well. This tends to happen with classes that exhibit a large intra-class variation in appearance, but could be easier defined through their function.

Also consider Fig. 1. Apart from large intra-class variation in appearance, it illustrates additional challenges that impede class detection. Scale and position of the objects play an important role. Some of the depicted objects might be categorized as "chairs", however one cannot really sit on them. The shadow on the wall still very much has the shape of a chair, but it should not be considered as such.

Thus, it stands to reason that adding affordance cues can help in resolving such cases. Especially if appearance is not what is mainly shared by class members, it makes sense adding such features. For instance, a dictionary[3] gives the following definition for a chair:

> "*chair*: a *seat* (→ something designed to support a person in a **sitting** position), esp. for one person, usually having four legs for support and a rest for the back [...]."

The *function*, *i.e.*, what one can *do* with the object, is more important to this definition than its form.

Of course, it will also rarely happen that affordance or function fully define an object class (*c.f.*, [9, p. 134]). For example, *sitting* can be done on chairs, stools, sofas, *etc*. Equally, the object can have more than one affordance. Fig. 2 illustrates the difference between functional and visual categories. Also, affordance depends on both the object and the actor. An adult will not sit comfortably on a child chair, and neither would a small kid on a chair for an adult. But again, it is function which can help to tell them apart, more than appearance.

Thus, in this paper, we propose affordance detection as additional, complementary cue for scene analysis. For its implementation, we will rely on 3d information. With the advent of cheap depth cameras, such data will become readily available.

[3] `http://www.dictionary.com`, 2010/10/02.

Related Work. Most recent object detection and categorization methods work on 2d images and follow the same principle. Image features are extracted and clustered like in the bag-of-word approach [6]. A single classifier or a set of classifiers is then trained on the features. Depending on the features and the classification techniques, a wide variety of different approaches have been proposed over the last years, *e.g.*, [24, 7, 15]. Some methods include more knowledge about the object and use 3d models (*e.g.*, [19, 14]) or build multi-view 2d models (*e.g.* [21]). However, all these approaches focus on the object appearance itself.

The concept of affordance [9, p. 127ff] has become a focus of attention within the cognitive vision and robotics community lately, *e.g.*, [25, 1]. Already in the early nineties, Stark and Bowyer [22] have proposed the use of functional properties. The 3d description of an object is parsed in the search of potential functional elements, which are then used to recognize the object. More recently, approaches have been proposed to detect objects based on human interaction. The human activity is annotated by extracting human motion from video data and used to indirectly identify objects [10, 13, 20]. In these works, it is assumed that interactions are observed during training and testing. Furthermore, Stark *et al*. [23] have proposed to learn from human demonstration. They use affordance as cue together with image segmentation to learn important object features while one interacts with the object. The detection is then performed with the selected appearance features.

In summary, these approaches have shown that function is an important property. They, however, either assume that the interaction with the object is also observed during detection, *i.e.*, one requires someone to interact with the object, or use affordance only as selection process for appearance features. In contrast, we propose to hallucinate the interaction during detection and to learn an *affordance detector*.

2. Affordance Detection

In this work, we assume that it is functionality one wants to retrieve. We are not so much interested in localizing instances of the particular object category "chair", but rather spots for "sitting". This might indeed be not only chairs but also a bench or a sofa. Affordance is handled as a cue complementary and parallel to appearance, rather than being a consideration after appearance-based detection. And, as our experiments show, affordance also significantly improves appearance based class detection, like the detection of "chairs".

2.1. Imagine Actor-Object Interaction

We model the affordance as interaction between a virtual actor and example objects. The proposed approach is illustrated in Fig. 3 for "sitting". In the figure, the chair affords the actor to sit on it. If one can imagine sitting comfortably

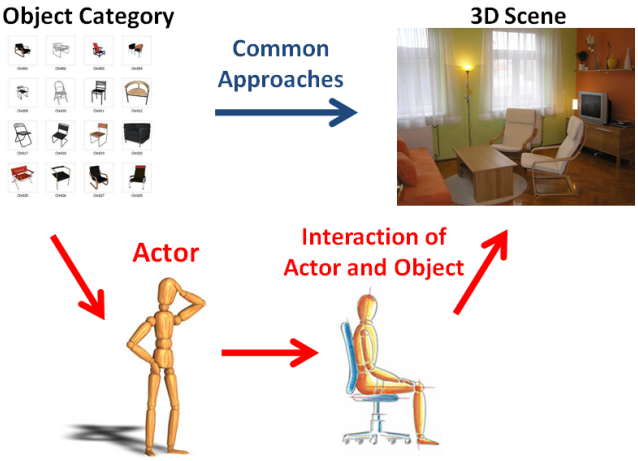


Figure 3. Common approaches for object detection use many training samples to build an appearance model of the category. However, many categories might be better described by the functions they support. Hence, we propose to imagine how well an actor could perform a specific activity with a scene part, in order to detect such objects.

at some spot in a scene, there is evidence for the existence of an object class with such affordance. We assume the availability of 3d data to probe the interaction, as coming from structure-from-motion or depth cameras.

In this paper, we focus on objects that involve full body human interaction. Furthermore, we only consider interactions where the action can be described by key poses. For example, a typical sitting position of a human infers a sittable place. However, the general principle can also be applied to body parts or other objects. In general, the concept of actor corresponds to the active part in an object-object relation, which can be a human as in our example, but it might be also another object, *e.g.*, a key that opens a lock.

Our approach is inspired by shape sorting toys as shown in Fig. 4*(a)*. To perform matching between shapes, we represent the human action by key poses that are matched with the 3d scene, see Fig. 4*(b)*. For instance, when observing a chair or a toilet seat, one can directly imagine how to sit on the objects. However, the traffic cone does not imply a comfortable sitting posture as illustrated in Fig. 4*(c)*. Using an actor for representing the functionality of an object has the advantages that (i) the relevant parts for the functionality are automatically recovered from the observed actor-object interactions and that (ii) the relevant parts are mapped to a unified representation, namely the actor. The core of the concept is a probabilistic model defined on the actor's shape that is able to encode both variations in action style and variations in object shape.

2.2. Model

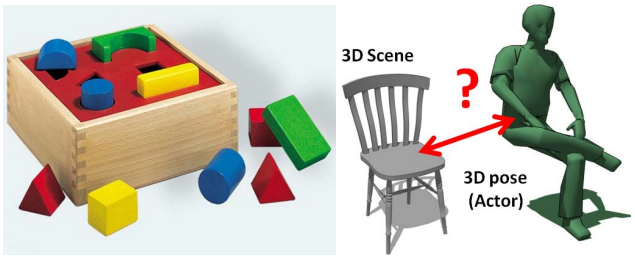
In order to learn the relation between a human and an object, we require at least one training example where we observe the functionality, *i.e.*, showing the object in use.

Key Poses. For each training example i, we assume a model of the object represented as 3d triangle mesh $\mathbf{M}_i^{object}$ and a model of the actor interacting with the object, which is also represented as 3d triangle mesh $\mathbf{M}_i^{actor}$. We further assume that the triangles and connectivity of the meshes $\mathbf{M}_i^{actor}$ are consistent over all training examples. This is achieved by using a consistent human model [11] for annotation.

To make the detection process efficient, we reduce the number of poses to a small set of key poses $\bar{\mathbf{M}}_k^{actor}$. The key poses can be obtained by clustering and taking the mean of each cluster, *i.e.*, the vertices $V_k^j \in \bar{\mathbf{M}}_k^{actor}$ are given by

$$V_k^j = \frac{1}{K} \sum_{i_k} V_{i_k}^j, \tag{1}$$

where K is the number of training examples i_k within cluster k and $V_{i_k}^j$ denotes the j-th vertex of mesh $\mathbf{M}_{i_k}^{actor}$. For simplicity, but without loss of generality, we now refer to a single key pose $\bar{\mathbf{M}}^{actor}$.



(a) Shape sorting toy (b) Actor-object matching

(c) Imagine sitting

Figure 4. Inspired by shape sorting toys *(a)*, we define the functionality detector as an actor-object matching problem *(b)*. One can imagine to sit on a chair or similar objects, but the traffic cone does not imply a comfortable sitting posture *(c)*.

Affordance model. For modeling the relation between the key pose and the objects $\mathbf{M}_i^{object}$, we rely on simple geometric features, namely 3d distance and mesh intersections between the actor and the object.

For the 3d distance, we voxelize the mesh $\mathbf{M}_i^{object}$ and compute a 3d distance field $\mathbf{D}_i$ [12] of the scene with the object. The closest distance of a vertex $V^j \in \bar{\mathbf{M}}^{actor}$ to the object i is then efficiently obtained by

$$d_i^j = \mathbf{D}_i(\mathbf{T}_i V^j), \tag{2}$$

where the rigid transformation $\mathbf{T}_i$ consisting of rotation and translation maps the coordinate system of the human model to the correct position inside the coordinate system of the object. In this way, we obtain a set of distance values for each vertex V^j and reconstruct the underlying probability using a kernel density estimator with a Gaussian kernel:

$$p_j^{dist}(d) = \frac{1}{n\sqrt{2\pi\sigma^2}} \sum_{i=1}^{n} \exp\Big(- \frac{(d - d_i^j)^2}{2\sigma^2} \Big), \tag{3}$$

where n is the number of training examples.

For the mesh intersections, we evaluate for each triangle $T^l \in \bar{\mathbf{M}}^{actor}$ whether it intersects with one of the triangles of $\mathbf{M}_i^{object}$. The intersection test is denoted by I_i^l, which is 1 in case of an intersection and 0 otherwise. Similar to the distance, we model the probability for a triangle T^l to have an intersection with an object, *i.e.*, $I = 1$:

$$p_l^{inter}(I) = \begin{cases} \frac{1}{n}\sum_i I_i^l & \text{if } I = 1, \\ 1 - \frac{1}{n}\sum_i I_i^l & \text{if } I = 0. \end{cases} \tag{4}$$

In the case of sitting, the probability of an intersection never exceeds 0.5.

Detection. The problem of detecting the functionality of a previously unobserved object $\mathbf{M}^{object}$ is formulated as a probability estimation problem, namely estimating the conditional probability $p(\mathbf{T}|\mathbf{M}^{object})$ defined over all transformations $\mathbf{T}$ of the key pose model $\bar{\mathbf{M}}^{actor}$ into the scene $\mathbf{M}^{object}$. Hence using Eqs. (3) and (4), our model becomes

$$p(\mathbf{T}|\mathbf{M}^{object}) \propto \tag{5}$$
$$\Bigg(\prod_{j=1}^{|V|} p_j^{dist}(\mathbf{D}(\mathbf{T}V^j)) \Bigg)^{\frac{1}{|V|}} \cdot \Bigg(\prod_{l=1}^{|T|} p_l^{inter}(I_{\mathbf{T}}(T^l)) \Bigg)^{\frac{1}{|T|}},$$

where $I_{\mathbf{T}}(T^l) = 1$ if the triangle T^l intersects with a triangle of the scene after applying $\mathbf{T}$ to the mesh; otherwise, $I_{\mathbf{T}}(T^l) = 0$.

For localizing objects or places that share the same functionality, we do not assume dynamic content or data where a human is part of the scene. Instead, we hallucinate the human interacting with the scene, *i.e.*, we densely evaluate $p(\mathbf{T}|\mathbf{M}^{object})$. To this end, we voxelize the scene and compute a 3d distance field. In our experiments, the transformation matrices $\mathbf{T}$ are parameterized by the translation vector (t_x, t_y, t_z) and the rotation θ around the axis perpendicular to the ground plane. Since we are only interested in transformations with high probability, the evaluation can be performed efficiently using coarse-to-fine grid search and cascading.

3. Experimental Results

We train a "sittable" affordance detector based on objects from the category "chairs". Afterwards, we compare and discuss our approach on a synthetic dataset. Finally, detection results on realistic scenes obtained through SfM or a depth camera are shown.

3.1. Dataset and Implementation Details

Data. There are several sources for acquiring training data, *e.g.*, 3d models available from the Internet can be used as well as models reconstructed from video, still images, or depth data. We downloaded 110 3d models of chairs from Google 3d Warehouse[4]. We randomly split the set of chairs into two subsets of 50 and 60 samples each, which are used for training and testing, respectively. As negative test samples, we use 662 samples (all except chairs and sofas) from the SHREC'09-Dataset[5].

Training. We semi-automatically fitted a statistical human model [11] to 10 examples and averaged the poses to obtain a key pose of a sitting person. This key pose was then manually placed on the 3d models of the chairs. Fig. 5 shows our obtained model. We choose $\sigma = 5$ for the parameter in Eq. (3), but very similar results are obtained for different values[6].

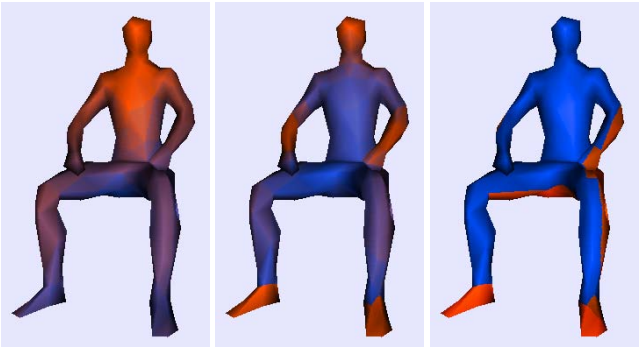
Detection. Voxelization is done with a voxel size of 1 cm^3. For evaluating Eq. (5), we perform a grid search starting with a stride of 8 voxels. The search is iteratively further refined to grid sizes of 4, 2, and 1 voxels. Points on the finer grid levels are only evaluated when they have at least one neighbor on a coarser level where the probability is above a given threshold. In our experiments, we set the threshold relatively low (0.0001) to get full recall for the coarse-to-fine grid search using a 6-voxel neighborhood. The rotation θ is evaluated for a set of discrete values at each grid level. We used 8 and 10 rotation values for the classification and detection experiments in Section 3.2, respectively.

Since evaluating the intersection term in Eq. (5) is more expensive than evaluating the distance term, we use cascading to reduce the computation time. We first compute the

[4] `http://sketchup.google.com/3dwarehouse/`, 2010/08/28.

[5] `http://www.itl.nist.gov/iad/vug/sharp/benchmark/shrecGeneric/data.html`, 2010/10/12.

[6] We performed all experiments with $\sigma = 1, 5, 10, 25, 50, 100$.



(a) Mean 3d distance (b) Var. 3d distance (c) Intersection

Figure 5. Illustration of the model (red: high, blue: low). *(a,b)* Mean and variance of the 3d distance of each vertex to the object. Regions with high variance correspond to variations among chairs, namely the size of the backrest (head), the existence of armrests (elbow), and the distance of the seat to the ground plane (feet). *(c)* Probability of an intersection of a triangle with an object.

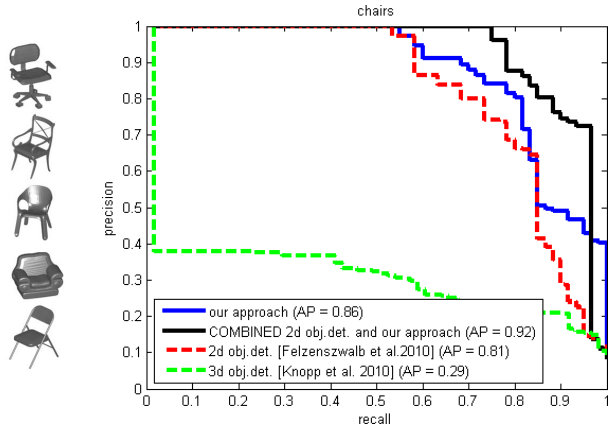


Figure 6. An affordance detector trained on the visual category "chairs". Our approach outperforms state-of-the art object detectors. The combination of appearance and our functional approach increases performance further. Hence, function is a complementary cue for object detection, rather than being considered after appearance-based detection.

probability based on the 3d distance field. When the probability is below a given threshold, the transformation $\mathbf{T}$ is rejected. Otherwise, we also compute the intersection term. In our experiments, we used the same threshold as for the coarse-to-fine grid search. For a scene with a chair that is discretized into 6×10^7 transformations $\mathbf{T}$, the approach requires about 15 seconds on a standard PC (single thread).

3.2. Synthetic Google-Chair Data Set

As basic experiment, we show the concept of the proposed affordance detector for classification and detection on synthetic data.

Comparison. We compare our approach to (i) a recently proposed 3d object classification system [14] and (ii) to a state-of-the art appearance detector working in the 2d image plane [7]. The 3d classification system was trained with the provided positive chairs[7]. For the 2d detector, we use the already trained version, publicly available on the authors' webpage[8].

Classification. We first evaluate the classification task: chair *vs*. non chair. Chairs are located upright on the ground plane. Negative samples are randomly rotated and scaled in order to fit the size of a chair. In order to apply the 2d appearance detector, we render 36 images by rotating the object around the vertical axis. For the non-chair objects, 36 randomly chosen views are rendered. The detector is applied on all the images separately and the maximum[9] response across all views is defined as the final score.

For a quantitative evaluation, we use recall-precision curves (RPC) and average precision (AP) [5]. Results are depicted in Fig. 6. Our functional approach achieves superior results over both other methods. The 3d approach has shown good results for object recognition. However, in our setting it suffers from the relative small training set (few very high ranked false positives causes the drop in precision). The pre-trained 2d detector (on a huge training set) has typical problems of being not aware of the 3d structure and the function of the object. We have to emphasize that, in contrast to [14, 7], our approach is not scale invariant since scale is an essential cue for functionality, see Fig. 1.

Combination of Appearance and Function. By equally weighting, we combine the scores obtained from the 2d appearance detector and our functional approach.[10] The AP increases from 0.81 and 0.86, respectively to 0.92, see Fig. 6. Hence, function can be seen as complementary cue for object detection which in fact increases performance by over 10% in terms of AP.

Generalization (function). In order to test how well the detector generalizes to the affordance category "sitting", we downloaded 60 3d models from Google Warehouse that are typically associated with sitting, *e.g.*, sofa, bench, or stool. The results demonstrate that function detection complements appearance cues quite well in this regard. While the affordance cue successfully generalizes towards these other object types, generalization by the 2d and 3d chair detectors is not effective, as depicted in Fig. 9. Since the appearance varies widely within the affordance category, see Fig. 2, the functional similarity is not well captured by appearance-based detectors.

[7] We gratefully thank Jan Knopp for applying their method on our data.

[8] Release 4, classifier `VOC2009/chair_final.mat`, `http://people.cs.uchicago.edu/~pff/latent-release4/`, 2010/08/28.

[9] Using mean or median yields similar results.

[10] Individual scores are mean and variance normalized.

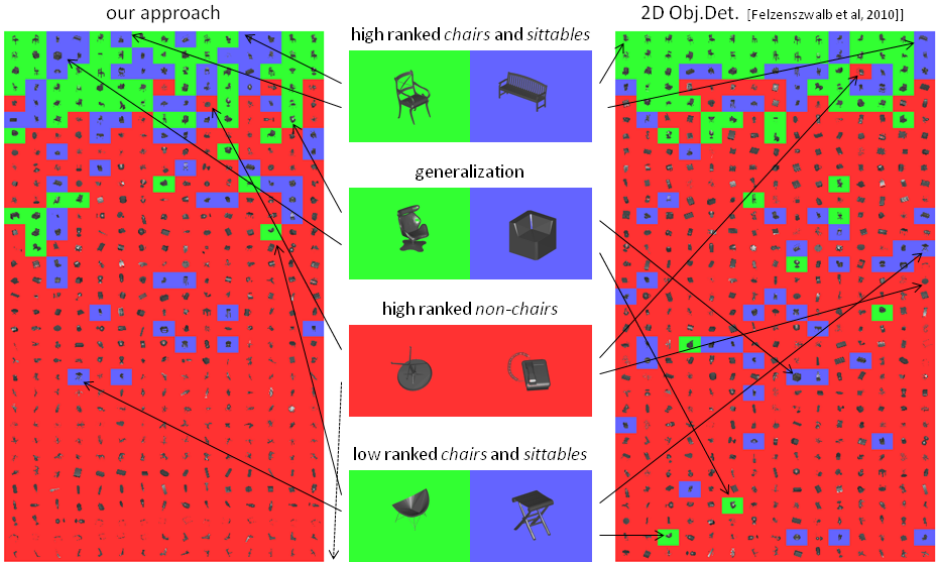
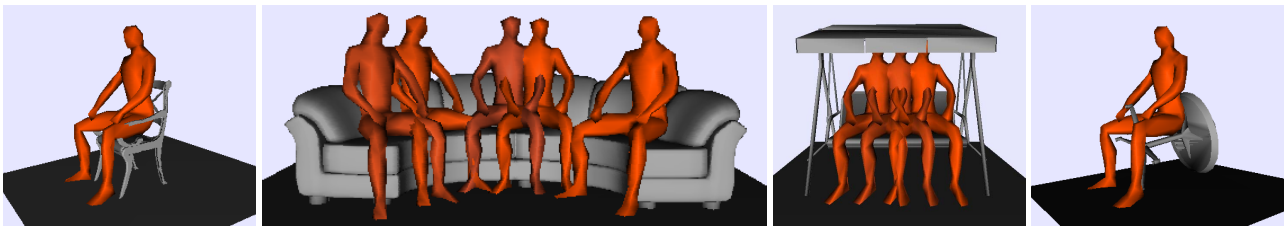


Figure 7. Ranking lists (top to bottom and left to right) of the 500 highest ranked objects obtained by our approach (left) and a state-of-the-art 2d appearance based object detector [7] (right). Both methods are trained on chairs only. Our method generalizes well to other chairs (green) and other sittable objects (blue) since it models rather the functionality than the appearance. Sittables that have a lower rank are stools without backrest since our model has learned the backrest as affordance cue from the chairs. [Full resolution figure can be obtained from the authors' web-page.]



(a) Chair (b) Sittable objects (c) High ranked "non-chair"

Figure 8. Few examples of detected sitting poses. Our approach also hallucinates a sitting pose for the table *(c)* since the tabletop is interpreted as backrest and the pole as something to sit on. Indeed, this is a valid sitting pose according to our model since our approach relies only on geometric properties and does not evaluate the physical stability of the sitting.

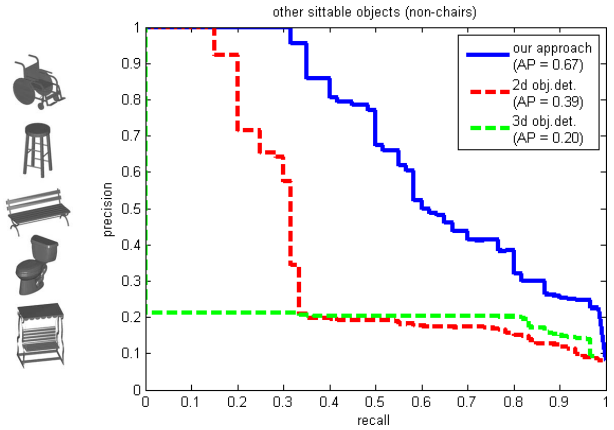
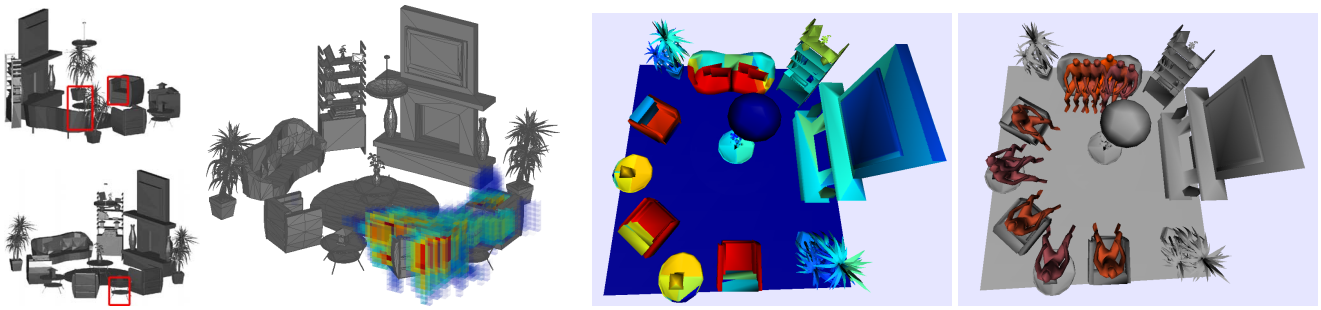


Figure 9. In contrast to appearance based detection our functional approach generalizes quite well to the affordance category "sitting". Hence, function complements appearance in this regard.

Interpretation. A more detailed comparison is shown in Fig. 7, where the 500 highest ranked test examples obtained by our approach (left) and by the 2d appearance based approach (right) are shown. Some objects appear with very high rank for both approaches. However, our approach trained from a few training samples generalizes very well across chairs and other sittable objects, whereas generalization for the 2d approach strongly depends on the appearance. For example, the bench is detected since the side view is very similar to a regular chair. Other sittable objects appear much lower in the ranking list. As can be seen from Fig. 7, our approach has mainly difficulties with stools. Since the detector has been trained on chairs, it assumes that the backrest is an essential part for sitting.

In contrast to object detectors, our method also predicts how to use the object by hallucinating the pose of the actor, as shown in Fig. 8. To measure the pose prediction accuracy, we have manually annotated the sitting poses for the test chairs. For the predicted transformations, we get 10.9 ± 11.0 cm error for translation and 0.53 ± 0.91 rad error for the orientation. The predicted poses can also be used to interpret the detections. For example, one can imagine sitting on the chair and the sofa, but it also seems possible to sit on the high ranked table in the given position, see Fig. 8*(c)*. Such reasoning is not possible for the 2d detector, which explains, for instance, the highly ranked phone.



(a) Sample detections (b) Combined scene view (c) Confidence map for sitting (d) Sitting poses with high probability

Figure 10. Detection sittable locations in a synthetic scene. The 2d appearance detector [7] casts votes according to its detections *(a)* in the 3d volume for detecting chairs. In contrast, our method yields a much more accurate confidence map *(c)*. The detected sitting places and corresponding poses *(d)* with high probabilities are very plausible, *i.e.*, high scores are obtained for the sofa and the chairs. The tables have lower scores, but still a relatively high probability for sitting.

Generalization (training set size). We reduce the training set size from 50 to 25 and to 10, without significantly decreasing the performance on the category "chairs", as well as on the functional category "sittable", see Tab. 1.

trainingset size	average precision chairs	sittable
50	0.86	0.67
25	0.87	0.66
10	0.85	0.63

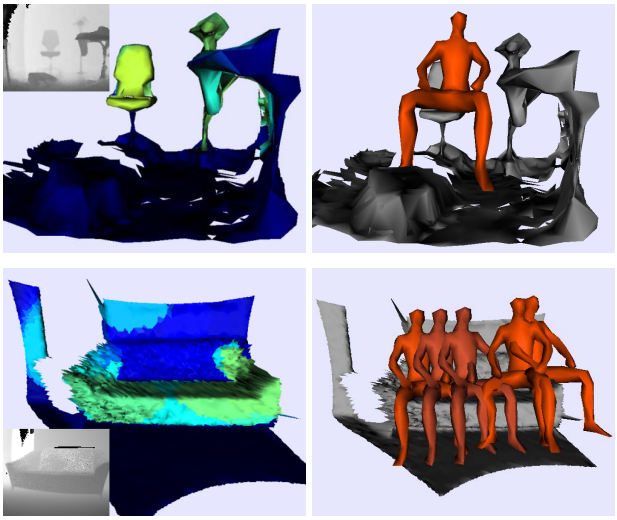
Table 1. Since our functional approach models only the relevant parts for sitting, the model can be trained quite accurately given only very few training samples.

Detection. Let us consider the task of detecting all sittable places in the synthetic scene shown in Fig 10. To make the comparison as fair as possible, we apply the 2d object detector on 36 rendered views from the scene. Due to the large intra-class variation of the objects, the cluttered background, occlusions, and pseudo structures due to the projection, one finds misaligned, missing, and false detections, see Fig 10*(a)*. The detections (low threshold) are then fused by casting votes in the 3d volume according to the camera rays and the detection score. Compared to the 2d detector, see Fig 10*(b)*, our approach shows a clear confidence map, see Fig 10*(c)*, with peaks at the three chairs and the sofa, but also yields some evidence for the small tables. The predicted sitting poses shown in Fig 10*(d)* are very plausible.

3.3. Real-world data

Finally, we applied the affordance detector trained on the synthetic chair data to real world scenes. In order to obtain the 3d scene structure we used to following two methods:

Depth camera. We used a time-of-flight camera to capture depth images from two scenes and applied our detector to the triangulated depth images. Results are shown in Fig. 11.

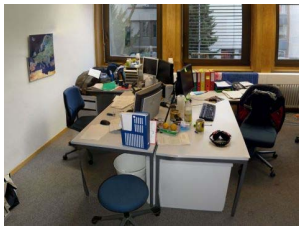


(c) Confidence map, depth image (d) Estimated sitting poses

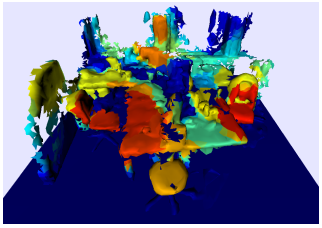
Figure 11. Two examples for data acquired with a depth camera. The sitting on the chair and on the sofa are well recovered despite the low resolution (176×144) and the noise of the sensor.

Structure from Motion. We recorded 80 images from a static office scene (see Fig. 12(a)). For dense 3d reconstruction, we use structure-from-motion[11] and the approach of Zach *et al*. [26]. This 3d model is then analyzed by our affordance detector. The resulting confidence map and potential sitting positions are depicted in Fig. 12*(b,c)*. The sitting poses look very plausible except for the pose on the monitor where the back is additionally supported by the wall. As in Fig. 8*(c)*, the detection can be explained by the scene geometry although the pose is physically unstable.

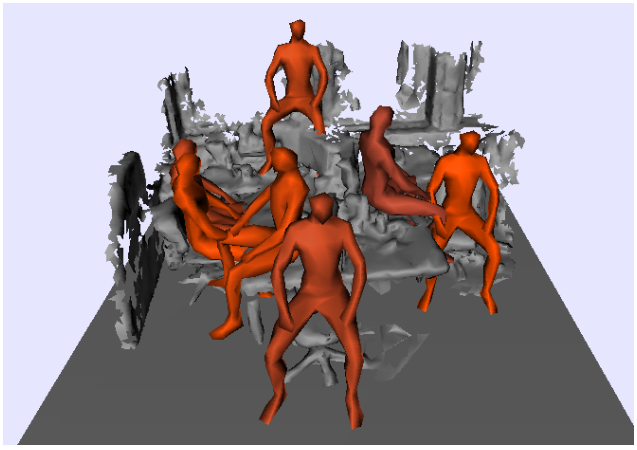
[11] `http://www.inf.ethz.ch/personal/chzach/opensource.html`



(a) Scene



(b) Confidence map



(c) Potential sitting poses

Figure 12. *(a)* An office scene is reconstructed from 80 images. *(b)* The sitting probability projected to the reconstructed surface. *(c)* Sitting poses with a very high probability. Besides the chairs, the stool and the table is recognized to have the functionality "sitting".

4. Conclusion

Objects are usually made for some purpose. Hence, the functionality often is the most obvious common denominator for the members of an object class. We have proposed an *affordance detector* where functionality is handled as a cue complementary to appearance, rather than being a consideration after appearance-based detection. By hallucinating an actor interacting with the scene, our approach predicts not only whether an object has the learned functionality, but also how the object can be used by the actor. We have demonstrated the potential of the method for the functional category "sitting", both for synthetic and real-world data. In both scenarios, objects or places that support this function are well localized and plausible sitting poses are recovered. Our current implementation relies only on geometric properties. Additional cues like physical stability or material properties are necessary to further improve the detection performance.

Acknowledgments. This research was supported by the EC Projects SCOVIS (FP7-ICT-216465), IURO (FP7-ICT-248314), and RADHAR (FP7-ICT-248873).

References

[1] E. E. Aksoy, A. Abramov, F. Wörgötter, and B. Dellen. Categorizing object-action relations from semantic scene graphs. In *Proc. Int. Conf. on Robotics and Automation*, 2010.

[2] A. Booth and S. Waxman. Object names and object functions serve as cues to categories for infants. *Developmental Psychology*, 38(6):948–957, 2002.

[3] I. Bülthoff and H. Bülthoff. Image-based recognition of biological motion, scenes, and objects. In *Analytic and holistic processes in the perception of faces, objects, and scenes*, pages 146–176. 2003.

[4] L. Carlson-Radvansky, E. Covey, and K. Lattanzi. "what" effects on "where": Functional influence on spatial relations. *Psychological Science*, 10(6):519–521, 1999.

[5] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge. *IJCV*, 88:303–338, 2010.

[6] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categorie. In *Proc. CVPR*, pages 524–531, 2005.

[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.

[8] E. Gibson. The concept of affordance in development: The renascence of functionalism. In *The concept of development: The Minnesota Symp. on Child Psychology*, volume 15, pages 55–81. 1982.

[9] J. Gibson. *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin Company, 1979.

[10] A. Gupta and L. Davis. Objects in action: An approach for combining action understanding and object perception. In *Proc. CVPR*, 2007.

[11] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A statistical model of human pose and body shape. In *Computer Graphics Forum (Proc. Eurographics 2008)*, volume 2, 2009.

[12] M. Jones, J. Baerentzen, and M. Sramek. 3d distance fields: a survey of techniques and applications. *IEEE Trans. on Vis. and Comp. Graphics*, 12(4):581–599, 2006.

[13] H. Kjellström, J. Romero, and D. Kragic. Visual object-action recognition: Inferring object affordances from human demonstration. *CVIU*, 2010.

[14] J. Knopp, M. Prasad, G. Willems, R. Timofte, and L. Van Gool. Hough transform and 3d surf for robust three dimensional classification. In *Proc. ECCV*, 2010.

[15] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 2007.

[16] M. Minsky. *The society of mind*. A Touchstone book. 1986.

[17] K. Nelson. Concept, word, and sentence: Interrelations in acquisition and development. *Psychological Review*, 81(4):267 – 285, 1974.

[18] L. Oakes and K. Madole. Function revisited: How infants construe functional features in their representation of objects. volume 36 of *Adv. in Child Development and Behavior*, pages 135 – 185. 2008.

[19] A. Patterson, P. Mordohai, and K. Daniilidis. Object detection from large-scale 3d datasets using bottom-up and top-down descriptors. In *Proc. ECCV*, 2008.

[20] P. Peursum, G. West, and S. Venkatesh. Combining image regions and human activity for indirect object recognition in indoor wide-angle views. In *Proc. ICCV*, pages 82 – 89, 2005.

[21] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *Proc. ICCV*, 2007.

[22] L. Stark and K. Bowyer. Achieving generalized object recognition through reasoning about association of function to structure. *PAMI*, 13:1097–1104, 1991.

[23] M. Stark, P. Lies, M. Zillich, J. Wyatt, and B. Schiele. Functional object class detection based on learned affordance cues. In *Proc. Int. Conf. on Computer Vision Systems*, 2008.

[24] P. Viola and M. Jones. Robust real-time object detection. *IJCV*, 57(2):137–154, 2002.

[25] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *Proc. CVPR*, 2010.

[26] C. Zach, T. Pock, and H. Bischof. A globally optimal algorithm for robust tv-l1 range image integration. In *Proc. ICCV*, 2007.