What makes a chair a chair?

By:
Helmut Grabner
Juergen Gall
Luc Van Gool

What makes a chair a chair?

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

What makes a chair a chair?

- Many object classes are primarily defined by their function
- The paper proposes a method to learn an affordance detector
- Imagine an actor performing an action typical to the target object class

Outline

- Introduction
- Modelling affordance
- Results
- Discussion

- «An object is first identified as having important relations, [...] perceptual analysis is derived of the functional concept [...]» - Nelson, 1974
- 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

Challenges impeding class detection

- Scale and position
- Shadows
- Something that can be sat on but is not a chair

- Affordance cues can help in resolving such cases.

A chair

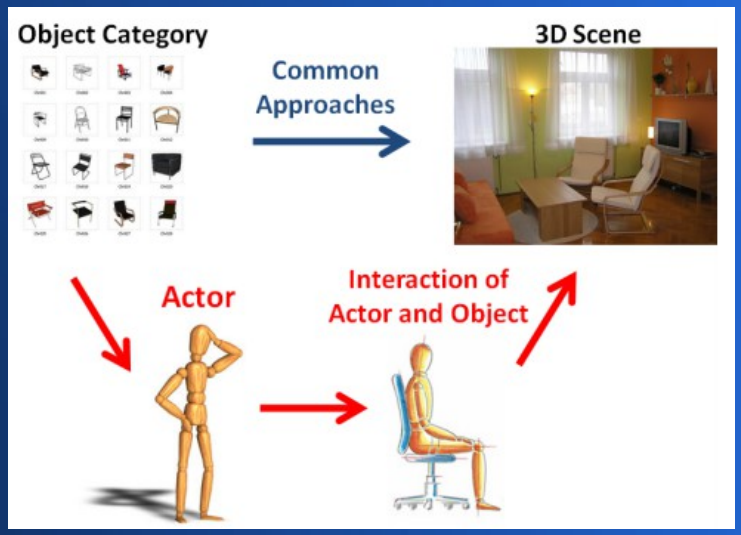
- «chair: a seat (-> something designed to support a person in a sitting position), esp. For one person, usually having four legs for support and a rest for the back»

In this paper

- An affordance detector
  - Limitations:
    - Only objects that involve full body human interaction
    - Only interactions that can be described by key poses
    - Requires 3D scene information

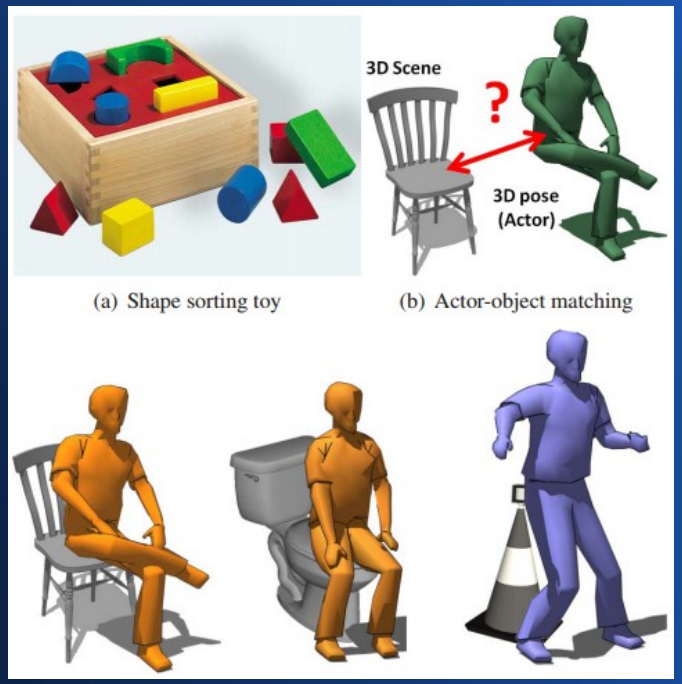
The model

- We are interrested in retrieving functionality.
  - Not a chair, but places to sit
- Model an interaction between a virtual actor and a target object:



The model

- Trains a classifier on a set of examples with key poses manually fitted to sample chair
-



Affordance model

- We rely on geometric features for modeling the relationship
  - 3D distance
  - Mesh intersections

3D distance

- Object is voxelized and a 3D distance field is computed
  - The closest distance of a vertex V^j to object i:

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

  - T is a transformation from the actor model to the object model

3D distance

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

- Reconstruct the underlying probability using a kernel density estimator with a Gaussian kernel
- n: number of training samples
-

Mesh intersections

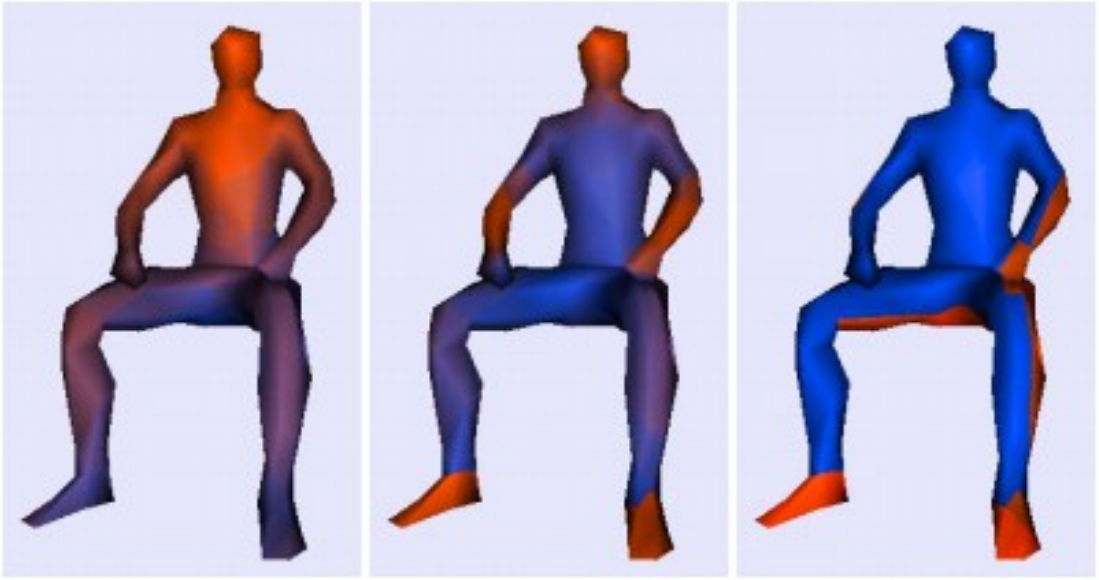
- For each triangle in the actor model, we evaluate whether it intersects with a triangle from the object model.
-

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

Evaluating

- Ends up being a probability estimation problem
- T are transformation matrices which are descretized on position and rotation (axis perpendicular to the ground plane).
- Mesh Intersection is only evaluated if p_j^{dist} is within a certain threshold

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



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

Training and testing (data)

- 110 3d models of chairs from Google 3D warehouse
  - Split into 50 for training, 60 for testing
- 662 negative samples (all except chairs and sofas)

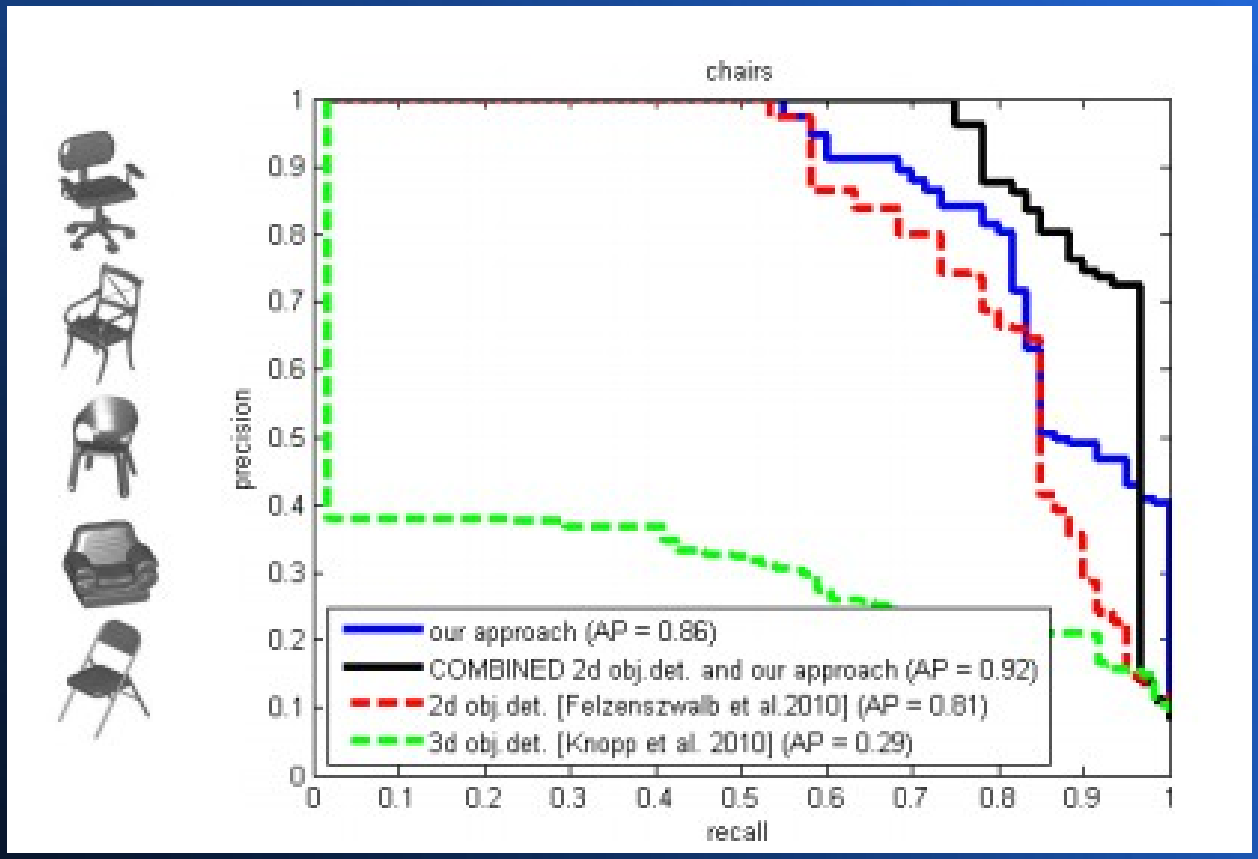
Training and testing (training)

- Human model semi-automatically fitted a human model to 10 key poses of a sitting person.
- Poses then averaged to find one key pose
- Key pose manually placed on 3D models of chairs.

Training and testing (detection)

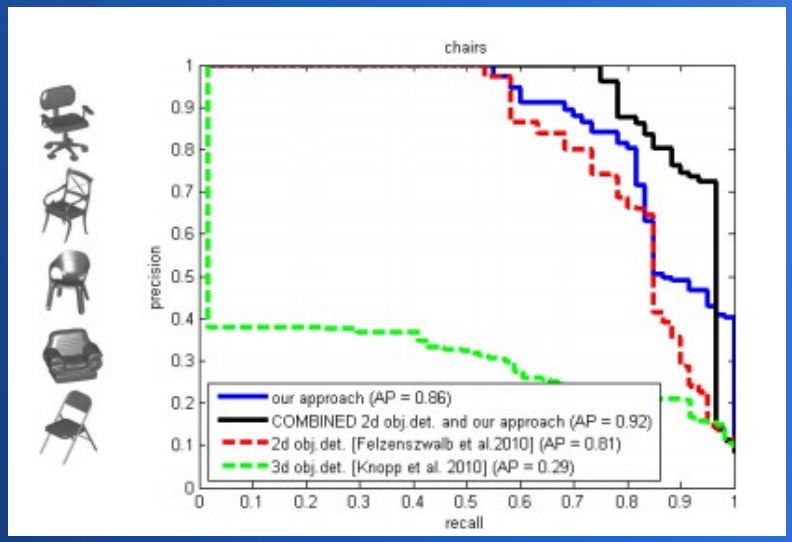
- Voxelizaion done with voxel size of $1cm^3$
- Grid search with stride of 8.
  - Refined to strides of 4, 2 and 1 when neighbours have a probability over a given threshold.

Results

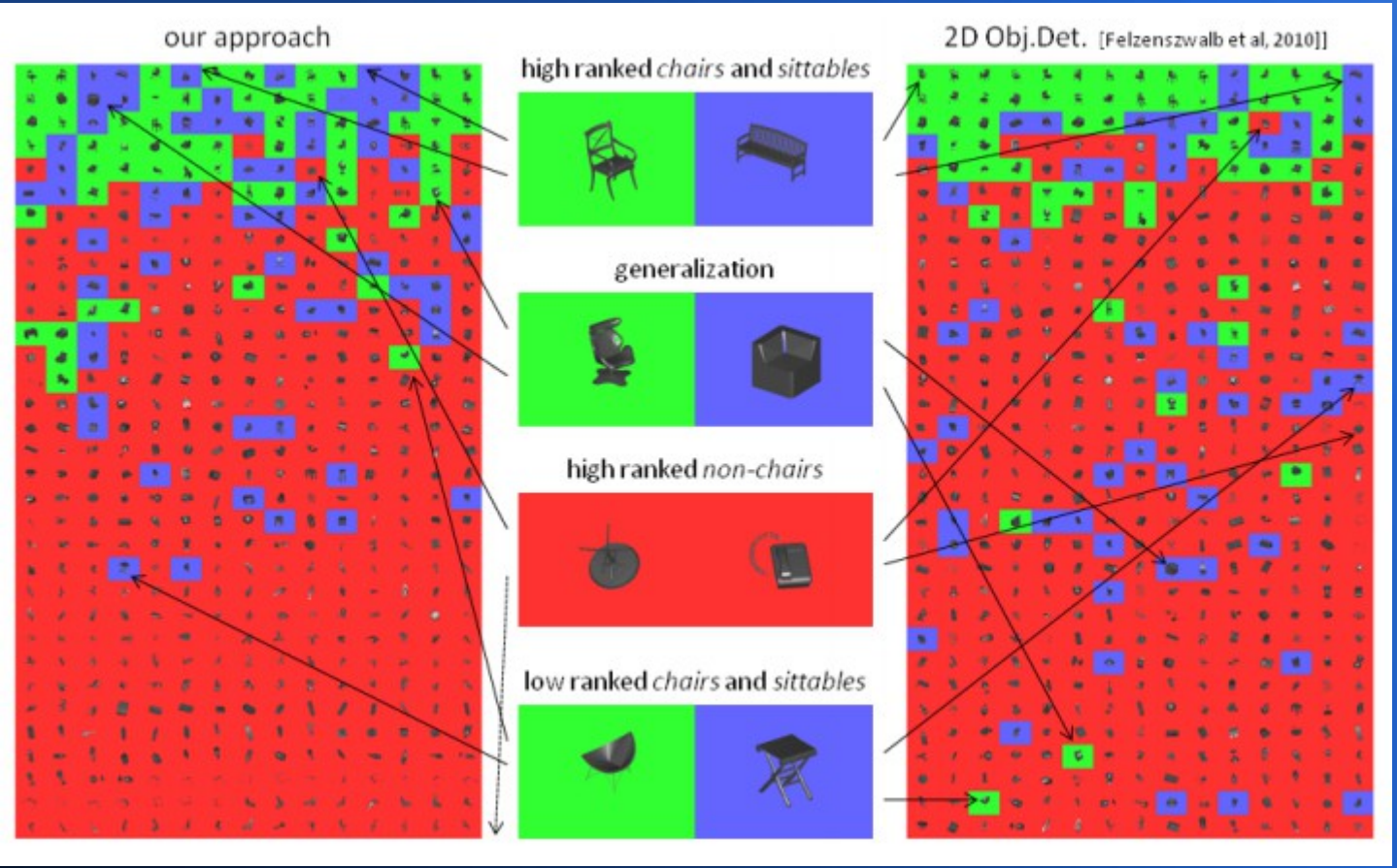

chairs
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
our approach (AP = 0.86)
COMBINED 2d obj.det. and our approach (AP = 0.92)
2d obj.det. [Felzenszwalb et al.2010] (AP = 0.61)
3d obj.det. [Knopp et al. 2010] (AP = 0.29)

Results

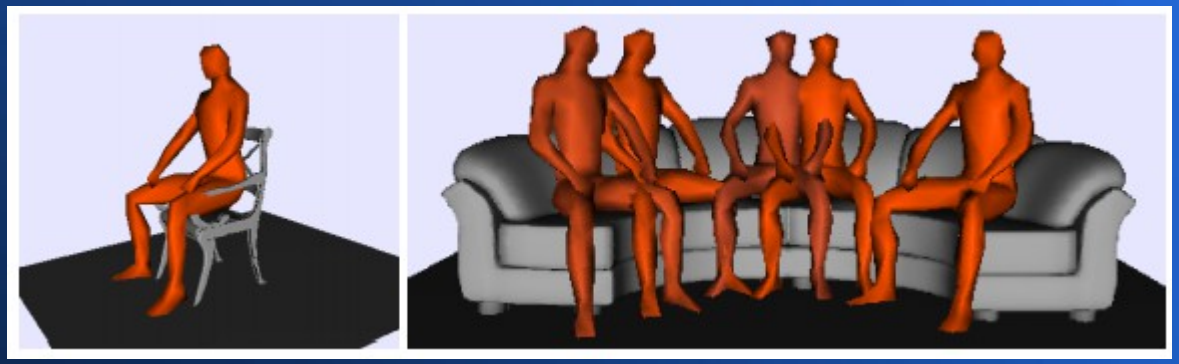
- Combining with 2D increases performance by over 10%
  - Affordance and appearance are complimentary cues

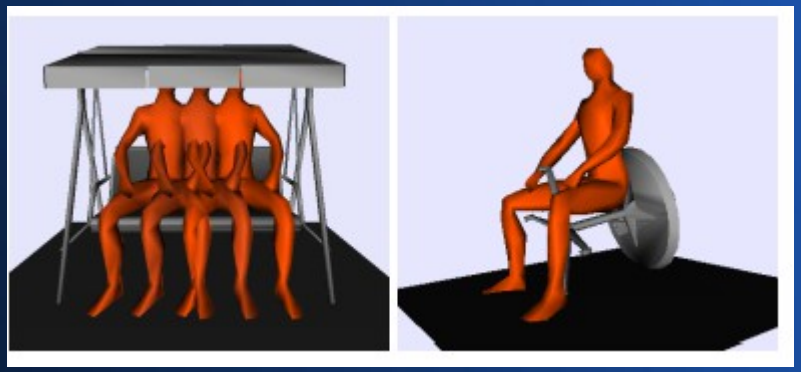


Results

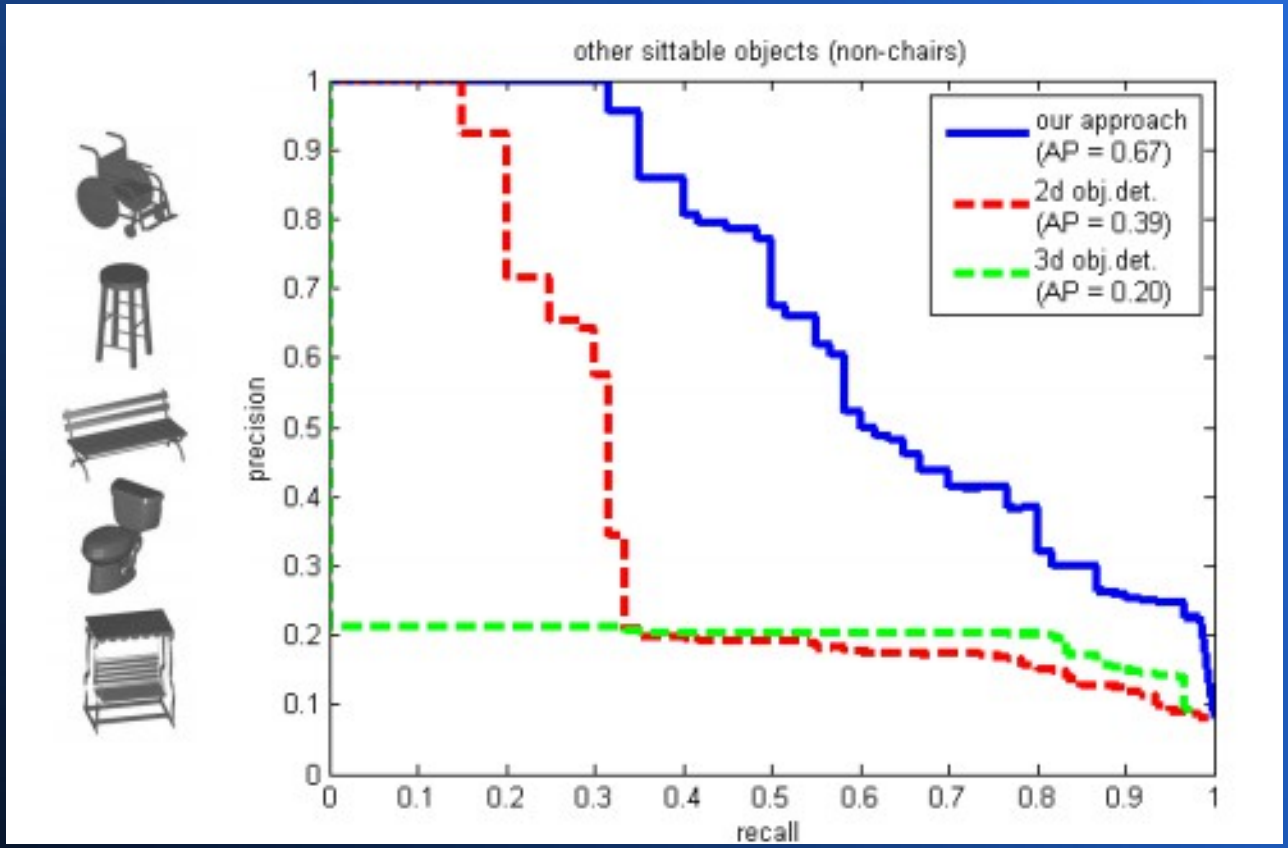

our approach
high ranked *chairs* and *sittables*
generalization
high ranked *non-chairs*
low ranked *chairs* and *sittables*
2D Obj.Det. [Felzenszwalb et al, 2010]]

Results

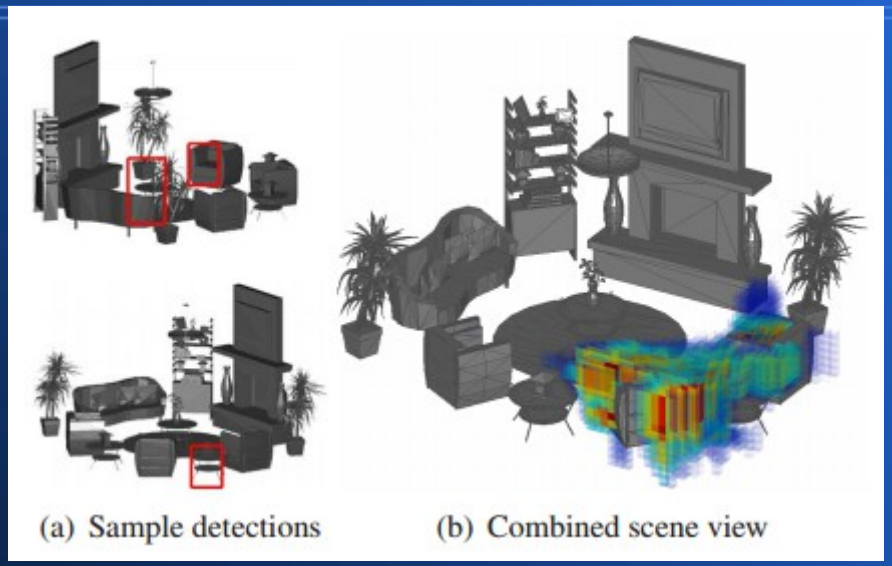




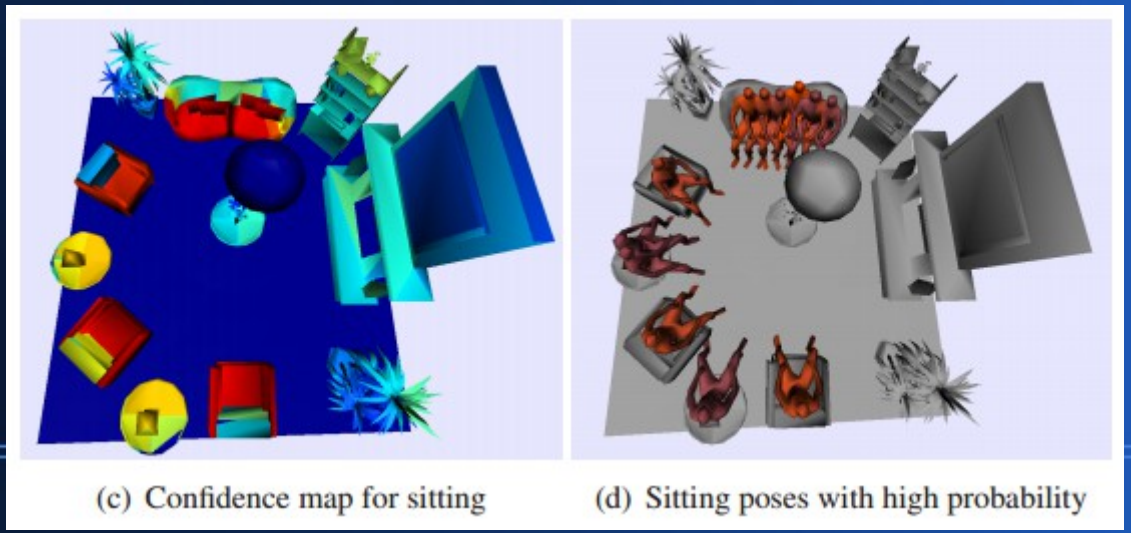
Results


other sittable objects (non-chairs)
our approach
(AP = 0.67)
2d obj.det.
(AP = 0.39)
3d obj.det.
(AP = 0.20)
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Results

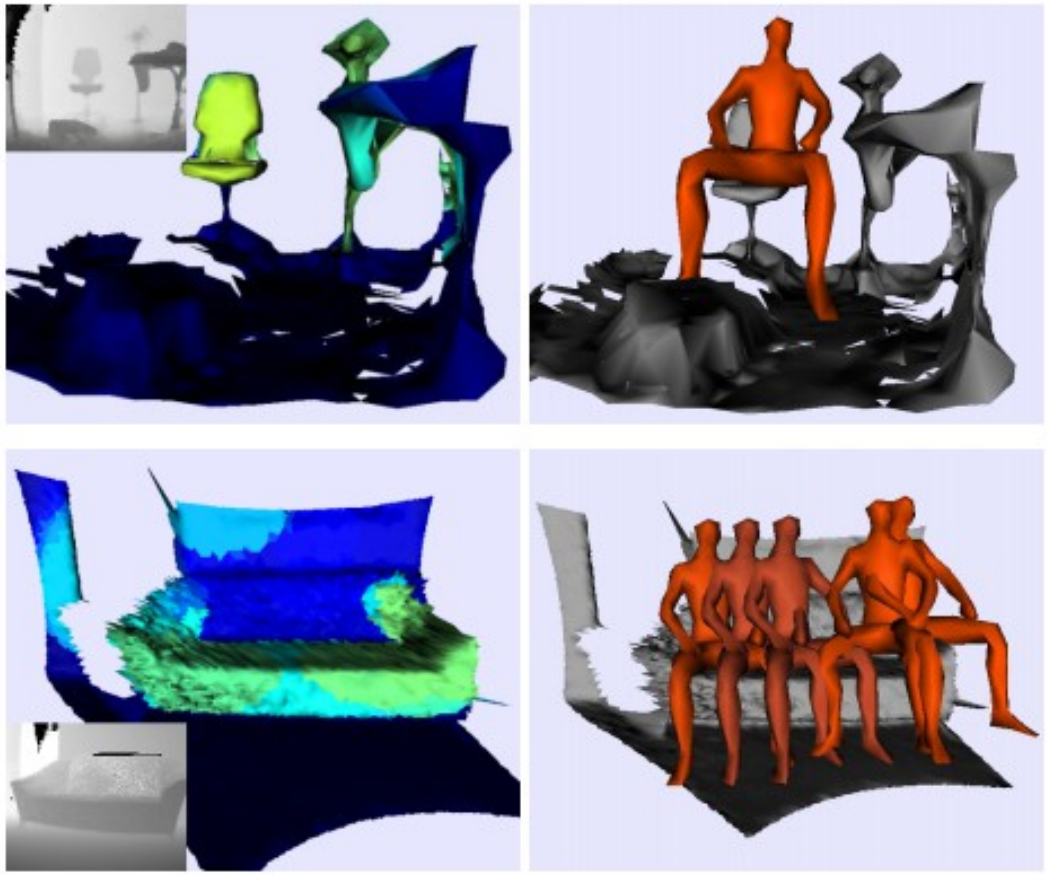


(a) Sample detections (b) Combined scene view



(c) Confidence map for sitting (d) Sitting poses with high probability

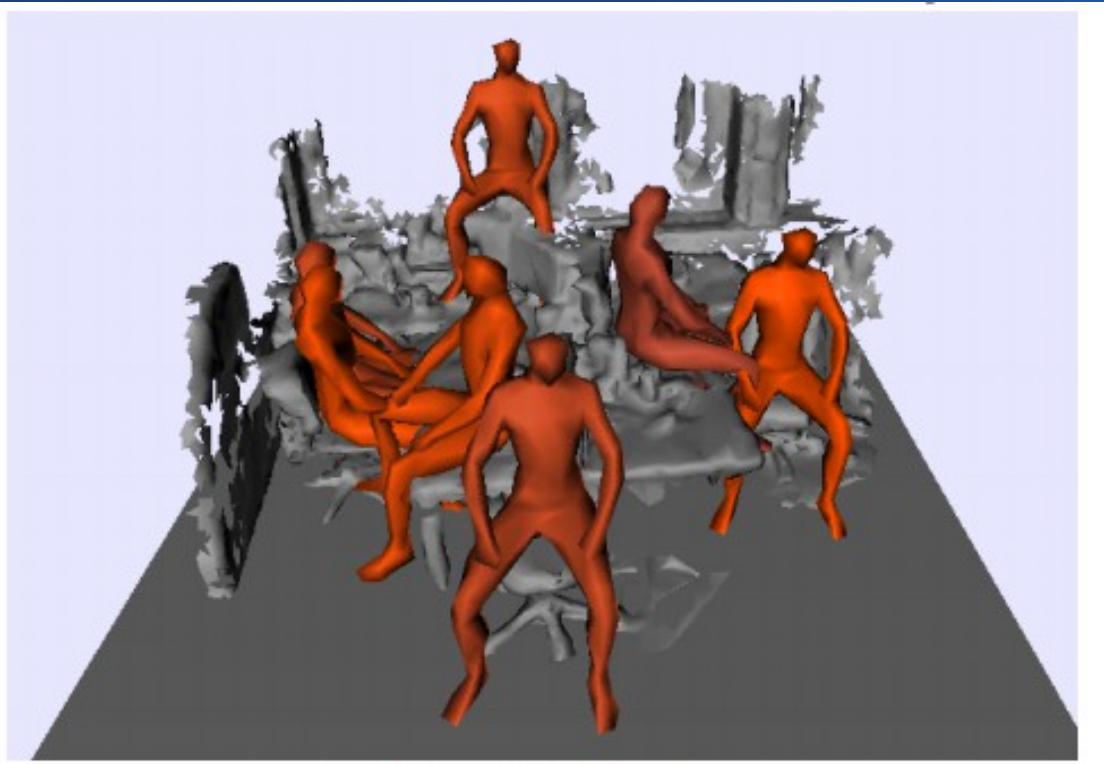
Results



(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.

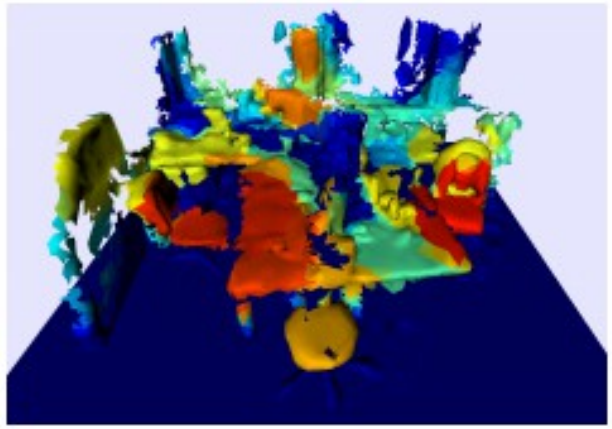
Results



(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".

(a) Scene



(b) Confidence map

Thank you for listening

Any questions?