From 3D Scene Geometry to Human Workspace

Abhinav Gupta, Scott Satkin, Alexei A. Efros and Martial Hebert
The Robotics Institute, Carnegie Mellon University
{abhinavg,ssatkin,efros,hebert}@ri.cmu.edu

Abstract

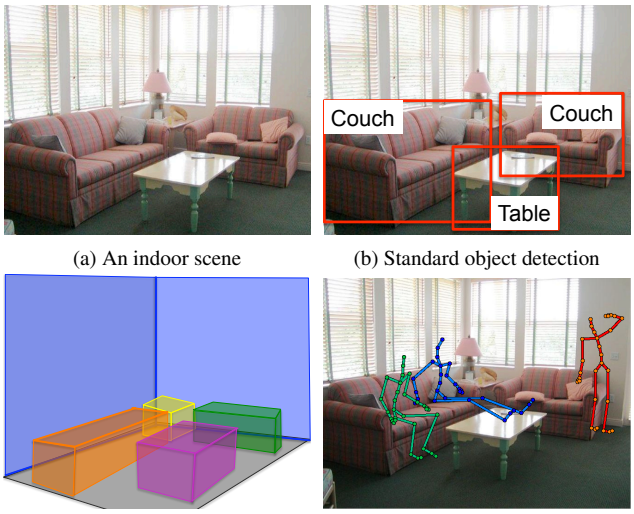
We present a human-centric paradigm for scene understanding. Our approach goes beyond estimating 3D scene geometry and predicts the "workspace" of a human which is represented by a data-driven vocabulary of human interactions. Our method builds upon the recent work in indoor scene understanding and the availability of motion capture data to create a joint space of human poses and scene geometry by modeling the physical interactions between the two. This joint space can then be used to predict potential human poses and joint locations from a single image. In a way, this work revisits the principle of Gibsonian affordances, reinterpreting it for the modern, data-driven era.

1. Introduction

Consider the image shown in Figure 1a. What would it mean to "understand" this image and how do we know if our understanding is reasonable and useful? This seemingly simple question actually hides a lot of complexity, going to the very heart of the computer vision problem. One popular answer is locating and naming the objects in the scene [6] (*e.g.,* "couch" and "table" in Figure 1b). However, understanding on the level of bounding boxes (or image segments) is rather superficial – it tells us little about where the objects are located within the 3D scene and how one can interact with them. While useful for various image retrieval tasks, such as searching for couch pictures on the Internet, a standard object detector would not help a blind man find where to sit.

To address these shortcomings, there has been a recent push towards more geometric approaches to image understanding [8, 11, 13, 16, 20, 22, 26]. The goal of these approaches is to recover an approximate, qualitative structure of the scene, typically modeled by planar surfaces or volumes (*e.g.,* Figure 1c). The advantage of such a representation is the ability to reason about the 3D space of the scene as well as the interactions (occlusions, depth ordering, proximity) between the objects within it.

However, what is often overlooked is that image *understanding*, unlike mere *measurement*, is a deeply subjective task. And the subject is *us*, the human observer. Implicitly, what we want from a computer vision algorithm is to understand our world *the way we do*. This means operating not



(a) An indoor scene (b) Standard object detection

(c) Geometry estimation (d) Our human-centric representation

Figure 1: There are many ways to "understand" an image; *e.g.,* by naming the objects within it (b), or by estimating the scene geometry (c). We present a human-centric scene representation that models the human workspace of a scene. It can predict human poses afforded by the geometry in a given image (d).

in terms of voxels or horizontal and vertical surfaces, but in terms of walkable space, sittable space, *etc.* We want vision to be human-centric, and justifiably so – for example, as flightless creatures, we *should* care more about an algorithm's correct estimate of the floor rather than the ceiling. Even semantic notions of object classes are usually impregnated with human-centric considerations: a "table" that is 6ft tall stops being a table, despite appearances, because its main functionality – reachability from the sitting position – is lost. At the same time, a low-standing filing cabinet could, under some circumstances, be considered a "chair."

The subject of this paper is putting the human back "into the picture" (Figure 1d). Our goal is to make indoor scene understanding more human-centric by reasoning about space from the point of view of a human actor. By analogy to the *workspace* of a robot, defined as the set of reachable poses given the geometry of a robot's mechanism [5], the idea is to represent a scene as a human workspace – that is, a set of reachable pose states for a typical human within the scene. The result is the intersection of objective scene geometry and subjective human pose. For

example, in Figure 1d, the "sitting reclined" action is the link that transforms the geometric blocks of the couch and the table into a functional chair.

Of course, this is a very difficult problem, and this paper is just the first step. The building blocks of this work, single-view indoor geometry estimation [11, 12, 15, 16] and human pose analysis via motion capture [3], are by no means perfect. However, we believe that now is the time to start designing ways to represent space and actions together, as it will guide further research in both geometry estimation as well as action recognition.

1.1. Background

The notion that objects are best defined not by their identity (e.g. "couch") but rather by their function (e.g. "sitting"), has a long and rich history. It dates back to early 20th century when Gestalt psychologists proposed that some functions of objects can be perceived directly. Kurt Koffka went so far as to claim: *"To primitive man, each object says what it is and what he ought to do with it: a fruit says, 'Eat me'; water says, 'Drink me', thunder says, 'Fear me' and woman says, 'Love Me'"* [14]. This idea was picked up and elaborated on by J.J. Gibson [7] who proposed the notion of affordances. Affordances can be seen as "opportunities for interactions" provided by the environment, which can be perceived directly from form and shape. While Gibson's affordance theory is very appealing, like Koffka's it may be going a bit too far. Do fruits really afford eating? The famous counter-example by Palmer [18] is "mailbox vs. trash can": since their shape can be very similar, it's impossible to infer their function just from their visible structure. In these cases, we use the association of an instance with past experiences to infer functionality. To address these shortcomings, Neisser [17] narrowed the concept and proposed physical affordances where only the *physical* interactions could be perceived from the physical structure of the objects. For example, while one can predict affordances like "throwable" and "pushable" from object structure, affordances like "for mail" cannot be predicted without associations.

In computer vision, over the last 30 years there have been regular attempts to use functional reasoning as a way to model objects by their "purpose" rather than their appearance [4, 19, 21, 23]. Most were recognition approaches that first estimated physical attributes/parts and then jointly reasoned about them to come up with an object hypothesis. For example, Stark *et al.* [21] used manually-defined rules to reason about functional elements of 3D CAD objects (*e.g.,* chairs) for recognition. A typical rule would define a particular type of chair as a combination of legs, surfaces of given heights and orientations, handles, *etc*.

Unfortunately, the hopes that these early efforts would generalize beyond the few hand-picked object classes with detailed 3D CAD models did not materialize. We believe that there were two major reasons for this:
(1) Use of Semantic Categories: While functional approaches aimed to move beyond semantic object categories into functional descriptions, these descriptions were, nonetheless, still semantic. For example, instead of defining the "chair" category, they would instead define the "sitting" category, which, while somewhat different, still shared all the problems of semantic object recognition.
(2) Reliance on Exact 3D: The early approaches were too optimistic in expecting to somehow obtain accurate segmentations and recover perfect 3D geometry from images – which decades later, still remain challenging problems.

In this paper, we present a physical model for capturing human-scene interactions[1]. Our approach uses 3D human poses as the functional categories and predicts the workspace of a human in terms of poses that can occur in a given scene. Furthermore, instead of assuming the availability of perfect 3D scene geometry, our approach builds upon the advances in 3D scene understanding from a single image and predicts the human workspace based on their geometric representation.

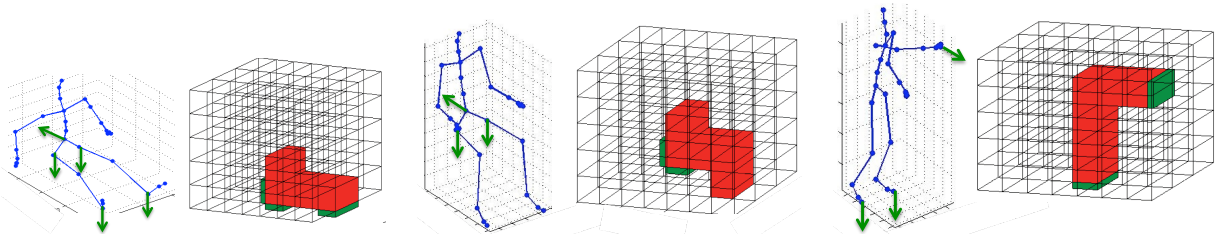
Our approach is most related to recent work in modeling geometric scene structure from a single image: inferring qualitative geometry of surfaces [13], grouping lines into surfaces [16], and estimating volumetric representations of a scene [8, 12, 15]. However, these approaches do not consider the interaction of humans with geometry. On the other hand, in recent years, motion capture technology [3] has allowed the routine capture of human form and shape. Researchers in computer graphics have looked into generating sets of constraints on human motions when interacting with objects in an environment [24]. However, in many cases, these constraints as well as the scene geometries are hand-defined, reducing the problem to path planning.

2. Overview

Our work is an attempt to marry 3D scene understanding with human action modeling. We propose a novel qualitative scene representation that combines 3D scene geometry with a set of possible human actions, to create a joint space of human-scene interactions. The ultimate goal is to be able to predict the full human workspace of the scene, that is the space of all physical actions that a given human agent can perform within a given scene.

In this paper, we present a proof-of-concept system for estimating a human workspace from a single image. We decided to limit our focus to indoor scenes, since they allow for more interesting human interactions, and since several approaches exist specifically for estimating indoor scene geometry [11, 15] (Section 3). We have selected a representative set of common physical human actions, such as reaching and sitting. While temporal information can provide additional constrains for many human actions, presently, we

[1]There have been a few recent papers which model the semantic relationships between humans and objects in 2D using appearances [9, 25]; however, our focus is on geometric and physical reasoning in 3D using form and structure.



(a) Sitting Reclined (b) Sitting Upright (c) Reaching

Figure 2: Qualitative Representation of Human Poses: Each pose is represented by the occupied blocks in discretized 3D space (shown in red) and the required surfaces of interaction (shown in green).

only deal with static poses. Therefore, we define the allowable actions by a set of 3D human poses, manually selected from a large motion capture database (Section 4). In Section 5, we formulate our human-scene interaction model and demonstrate it on synthetic 3D data. Then, in Section 6, we describe how to apply it to real images. Qualitative and quantitative results are presented in Section 7. Lastly, in Section 8, we discuss some potential applications of our new representation.

3. Representing Scene Geometry

Figure 3: Block world representation of an indoor scene. Each scene is represented by the walls (shown in red and blue) and the occupied voxels in the scene.

We adopt a geometric representation that is consistent with recent approaches for extracting indoor scene geometry from a single image [11, 12, 15, 16]. We briefly describe our scene geometry representation here: An indoor 3D scene is modeled by the layout of a room (walls, floor and ceiling) and the location and layout of the objects in the room. It is assumed that there are three principal directions in the 3D scene (Manhattan world) and that all walls and objects are aligned to those three principal directions. A simplified room is represented as a box being viewed from the inside, and therefore the layout can be encoded with only a few parameters (the locations of each visible wall). The objects in a scene are modeled by a set of occupied voxels. An example scene representation is shown in Figure 3. In this scene the extracted walls are shown in red and blue. The two beds in the image are represented by the set of occupied voxels in the 3D scene.

4. Qualitative Human Pose Representation

We now describe our basic representation of human interactions. Given the volumetric (voxel-based) representation of the scene, we would like to predict all the possible actions that are consistent with the scene geometry. We propose to use a vocabulary of potential actions represented in terms of human poses rather than semantic categories such as "sittable" or "touchable." However, the space of human pose configurations is combinatorial in nature and therefore using raw pose data is computationally infeasible. Furthermore, we would like our representation to generalize such that we do not need to see all possible poses before predicting them. For example, we should be able to generalize from the pose of a person sitting on a chair (of a specific height) and predict the pose affordance of a couch (of a different height).

A key insight is to note that there are only two constraints on a 3D human pose that are relevant for embedding it within a given geometry: 1) the 3D space (volume) the pose occupies, and 2) the surfaces it is in contact with. To allow for generalization, we use a discretized representation of the 3D space occupied by the human actor. We divide the space around the human actor into blocks (Figure 2) and associate each block with a 0 or 1 based on whether the block is occupied or not. In addition, each block may require an external support in a particular direction. For example, in the sitting pose (with back support), we need a horizontal surface below the pelvic joint to support the body and a vertical surface to rest the spine (Figure 2b). In a similar manner, for the "reaching" action (Figure 2c) a horizontal support is required at the feet and a vertical surface of interaction is required to represent the point of contact of the hands. Since it is impossible to enumerate all potential directions of support, we make the simplifying assumption that the scene and object layouts are aligned with three orthogonal vanishing points.

We use 3D human poses from motion capture data [1] for our experiments. We manually associate each action with an exemplar pose, and annotate the required support and interaction surfaces. Given the mocap poses and their corresponding support annotations, we obtain a discretized representation of these poses as explained above. For our experiments, we use three, five and seven blocks in the x, y, and z directions respectively. Figure 2 shows a sampling of the discretized poses used in this paper. The arrows on the stick figure indicate locations where support or interaction surfaces are required, and the green blocks on the right show the position of these surfaces.

5. Human-Scene Interactions

In the previous sections, we have described a model of scene geometry, as well as a simplified representation of human poses. Using these poses and scene geometry as inputs, we can now ask the question, "Where in the scene can a human perform these actions?"

In order for a pose to be valid at a specific location, two constraints must be satisfied:

Free space constraint: The volume occupied by a human cannot intersect any objects.

Support constraint: There must be object surfaces in the scene which provide sufficient support so that the pose is physically stable.

Consider the pose "sitting." The support constraint states that there must exist a horizontal surface beneath the pelvis (such as a chair). The free space constraint ensures that no object prevents a person from sitting on the chair.

By discretizing the scene geometry into an occupancy matrix, we can efficiently search for poses which satisfy the free space and support constraints. We begin by creating a binary representation of the environment, where each cell of a 3D matrix is 0 if there is free space and 1 if occupied by walls, furniture, *etc.* Human poses are also discretized into a binary occupancy matrix using the same cell size as the environment (we chose a cell size of 3x3x3 inches for our experimentation). Now, we can simply perform a 3D correlation operation to compute the set of valid locations for a pose. A non-zero correlation at a given location indicates a violation of the free space constraint. Conversely, where the correlation is zero, the intersection of the cells occupied by the person and the environment is empty – thus, the pose is valid.

In a similar manner, we can compute whether each location in an environment satisfies a pose's support constraints. We create a set of *interaction blocks* which indicate locations in the environment where objects or support surfaces must be present (shown in green in Figure 2). Again, we use a 3D correlation operation to compute the set of locations with the correct geometry to afford the pose. Unlike the free space constraint, now we are trying to maximize the correlation score, to find locations where there are environment blocks present which align with the interaction blocks. A correlation score equal to the number of non-zero interaction blocks indicates that all support constraints are satisfied for a pose at that location. We take the intersection of valid support locations and valid free space locations to determine all positions in the environment which afford the pose.

This view of the interaction between human pose and scene geometry is a bit idealized. For example, in the real-world free space and support of the environment actually deform to allow a wide range of poses (*e.g.*, cushions of furniture deform to the shapes of our bodies). Additionally, when presented with a rigid object, the human body can adjust. For example, we often slouch on seats where the back support would otherwise be uncomfortable. To represent these interactions, we need to the relax the free space and support constraint a bit. We implement this by applying a small amount of erosion or dilation to the occupancy map of the environment prior to performing each correlation operation. More precisely, a small amount of erosion in the occupancy matrix allows configurations which would otherwise violate free space constraints to become permissible. Conversely, dilation of the environment allows locations which do not have the necessary support surfaces to grow to accommodate a pose. For example, if the seat of a chair is too low to comfortably sit on, dilation will raise the support surface, making the pose possible. Thus, we first erode a scene's occupancy matrix before performing the correlation to validate free space constraints, and dilate the matrix before performing the correlation to validate support surfaces and surfaces of interaction.

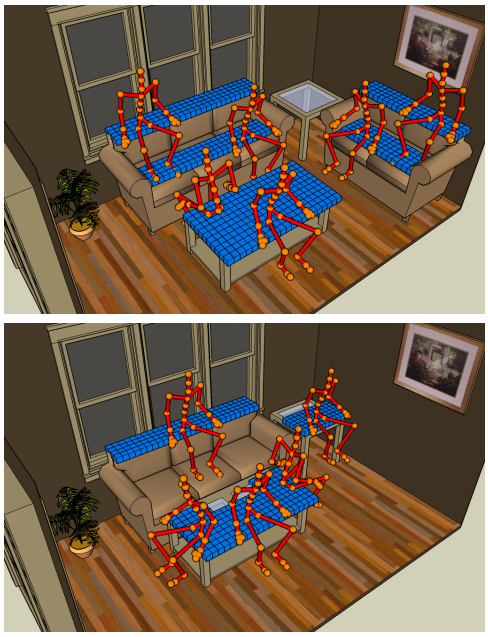


Figure 4: Human-centric representation on a synthetic 3D scene: given an example "sitting" pose, we visualize where a human can sit in the environment (blue mask shows all possible pelvis locations). Note how rearranging a few objects within the scene can have a big influence on the estimated human workspace.

We demonstrate the capabilities of our human-centric representation by running our human-scene interaction model on a few synthetic 3D indoor scenes downloaded from the Google 3D Warehouse [2]. Figure 4 shows the locations where a human can sit for two possible furniture configurations. It can be seen how our representation captures the spatial arrangements of objects and how affordances change for the same objects under varying configurations. While in Figure 4(top) a person can sit on the left couch, the same couch is no longer accessible for sitting in Figure 4(bottom) because the table is moved too close to it. Similarly, the side table becomes accessible in the bottom figure, once an obstruction is removed.

6. Geometry Estimation

We have shown how, given a scene geometry and a qualitative representation of human poses, we can derive the joint space of human-scene interactions using free space and support constraints. If the estimated geometry from a single image were exact, we could estimate the human workspace perfectly, presuming that the physical constraints provide a complete model of interaction (as demonstrated using ground-truth geometry in Section 5). However, estimating scene geometry from a single image is an extremely difficult problem. For robustness, we use two sources of geometry based on the outputs of [11, 15]. Given an input image, we select which source to use based on a cost function which measures the agreement between the geometry (placement of cuboids) and the estimated distribution of occupied voxels. An overview of our approach is shown in Figure 5.

Given an image (example shown in Figure 5a), we first compute a room layout hypothesis using the method of Lee *et al.* [15] (Figure 5c). We also compute the probability of each pixel being associated with object/clutter using the surface layout algorithm [11] (Figure 5b). The clutter labels are used to estimate an occupancy grid in 3D. A voxel in 3D is occupied by an object if: (1) the pixel corresponding to voxel center in image is classified as object/clutter, and (2) the voxel corresponding to the projection on the ground is occupied as well since most objects are supported by the floor. Each voxel in 3D space is projected onto the image using the standard projective camera model, where camera calibration[2] is performed using estimated vanishing points [10]. Therefore, the probability of each voxel ($V(X,Y,Z)$ where (X,Y,Z) is the location of the center of voxel) being occupied is computed directly from the probabilistic clutter image C as:

$$P(V(X,Y,Z)) = C(\mathcal{M}[XYZ1]^\top)C(\mathcal{M}[X0Z1]^\top) \quad (1)$$

where $\mathcal{M}$ is the camera projection matrix and $C(x,y)$ corresponds to the probability that pixel (x,y) belongs to clutter. Therefore, for a voxel to be labeled as clutter, both the pixels corresponding to the voxel and the pixels corresponding to the vertical projection on ground should be labeled as clutter. Figure 5d shows occupied voxels with probability $P(V(X,Y,Z)) > 0.5$. Once we have a probabilistic estimate of the occupancy of each voxel, our goal is to estimate the cuboids corresponding to objects that can explain the 3D voxel occupancy map. We generate two possible sets of cuboids (described below), and select the better set using a cost function consisting of two terms. The first term counts the non-occupied voxels in the cuboids and the second term sums the occupied voxels that have not been explained by any cuboids. The first terms can be written as:

$$f(S) = \sum_{O \in S} \sum_{V_i \in O} \log P(\neg V_i), \quad (2)$$

where S is the set of cuboids, O is a cuboid in that set, $V_i \in O$ is the set of voxels contained in O and $P(\neg V_i)$ is the probability that V_i is empty. The second term of our cost function can be written as:

$$g(S) = \sum_{V_j \notin \bigcup_{O_i \in S} O_i} \log P(V_j), \quad (3)$$

where V_j is the set of voxels that are not in any of the cuboids in the set S and $P(V_j)$ is the probability that V_j is non-empty.

This cost function is used to choose the better of two geometric hypotheses. The first hypothesis is obtained directly using the results of Lee *et al.* [15]. Here, the cuboid hypothesis for objects are generated by combining regions whose orientations are obtained using a sweeping algorithm. The final set of cuboids are chosen based on volumetric relationships and a learned cost function as explained in [15]. We generate the second hypothesis from the probabilistic 3D occupancy map (Figure 5d) by searching for standard size cuboids (seen frequently in training data and corresponding to common house-hold objects) that can explain the 3D voxel occupancy map. Under the assumption that most objects in a room rest on the ground and are attached to walls, we slide cuboids of multiple sizes (*e.g.,* 60x60x15in and 72x24x15in) along the walls of the room. We use a greedy approach which iteratively adds cuboids one by one such

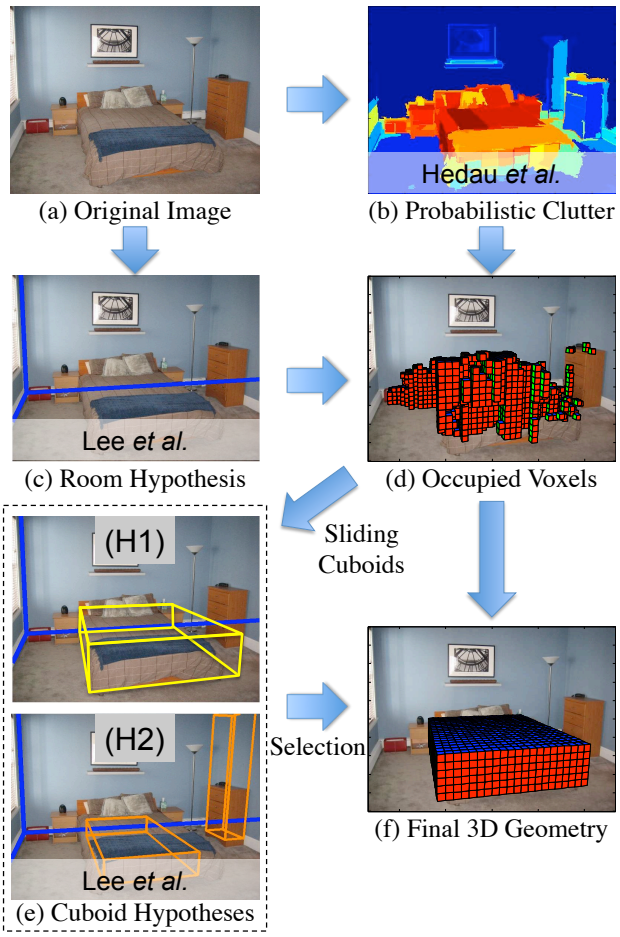


Figure 5: Overview of 3D Geometry Estimation.

[2]The axes of the world coordinate system are aligned with the principal directions (defined by the estimated vanishing points). The y-axis corresponds to the vertical direction. The projection of camera center on the ground is the origin.

that: (1) the cuboid should not intersect with other already selected cuboids, and (2) the cuboid has a high occupancy score. This can written as:

$$\arg\max_{O} \sum_{V_i \in O} \log P(V_i) \quad \text{s.t.} \quad \forall O_i \in S: \; O \cap O_i = \varnothing. \quad (4)$$

For robustness, we also impose the constraint that largest sized cuboids do not co-occur with smaller cuboids.

7. Evaluation

There has been little research in the area of human-centric scene understanding and therefore, there are no established datasets, methodologies or relevant previous work to compare against. We will present our experiments in two parts: 1) qualitatively, by showing a few representative results; 2) quantitatively, by comparing the performance of our approach to a baseline appearance based classifier for predicting the location of joints in different poses.

Tasks and 3D Poses: We show our results for four physical tasks: Sitting upright (no back support), Sitting reclined (back support and legs up), Laying down, and Reaching for a vertical surface. We manually select one pose for each task, except for the last one. Since the heights of the reaching locations can vary, we use four poses corresponding to the reaching task.

Dataset: We use the indoor scene dataset introduced by Hedau *et al.* [11]. The dataset consists of 314 images (209 training and 105 test images). Since we build upon the results of [11, 15], we use the same set of test images used in these papers.

7.1. Qualitative Evaluation

Figure 6 (next page) shows the performance of our approach on a representative set of images for which the automatic calibration procedure has low errors. To visualize the whole range of possible poses, we overlay colored masks indicating the locations of pertinent joints for a given pose. For example, we show in blue the locations where the pelvic joint makes contact with a valid surface of support for the "sitting reclined" task. We also indicate in cyan the locations where the back makes contact with a vertical support. Example human stick figures (extracted from the mocap data) show representative valid poses in each scene. As is evident from the stick figures, our approach predicts affordances that cannot be represented by basic object categories. For example, on the "sitting reclined" pose, our approach combines the vertical surface of the bed with the horizontal surface of the ground to predict human poses. Similarly, for "sitting upright" our approach finds valid pose locations that cannot be predicted by object-level categories such as chairs or couches. For example, in the second scene, our approach finds a table as one of the valid sitting locations and in a kitchen (fourth scene) it predicts the stove as a possible location for sitting. For the pose "laying down," our approach predicts beds, couches and the ground all as valid locations. Figure 7 shows a few failure cases when there is error in geometry estimation. For example, in the two-bed example, only one of the beds is detected. However, our approach shows graceful degradation and still predicts some correct locations for all poses. Subjectively, our approach is able to predict reasonable workspaces for about one-third of the images in the dataset, with the vast majority of errors coming from misestimated geometry. Our code and data is available online.

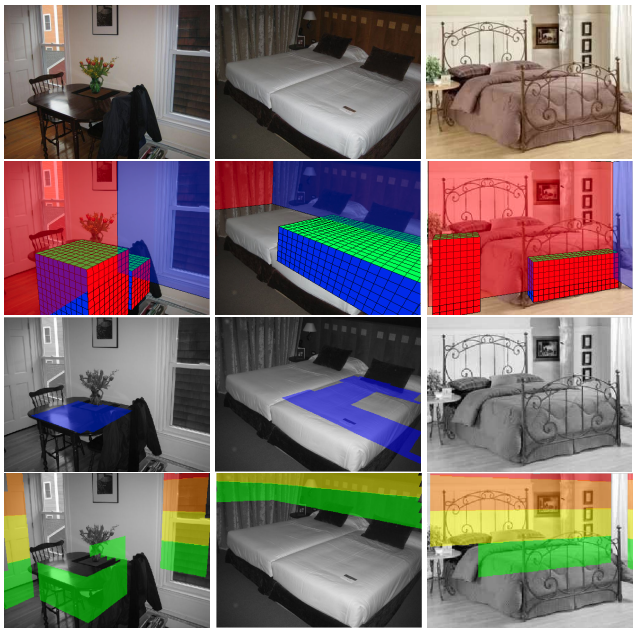


Figure 7: Additional results to show the performance of our approach when there is error in 3D geometry estimation. The 3D geometry estimation error increases from left to right (and the accuracy of our results drop accordingly). For example, only part of the bed is recovered in the second scene, and in the third scene, the estimated room layout is incorrect.

7.2. Quantitative Evaluation

Our approach allows us to predict possible human joint locations in a scene, which we now evaluate quantitatively. We compare the masks predicted by our algorithm to ground truth masks. We manually labeled 25 test images[3] for four poses:
(a) Locations a pelvic joint can rest while sitting upright,
(b) Locations a pelvic joint can rest while sitting reclined,
(c) Locations a human's back can rest when laying down,
(d) Locations a hand can reach on a vertical surface.

We also compare our algorithm with a standard appearance-based baseline; training a separate classifier for each task. These methods have shown good performance for different pixel labeling tasks, such as object categorization and qualitative geometry estimation. Each pose classifier uses appearance features computed from an image to label the pixels where a relevant body joint can appear for that human pose. For example, the "sitting upright" classifier predicts where a person can sit by indicating where the pelvic joint could rest in an image when the person is sit-

[3] We used images for which [11, 15] report reasonable vanishing point estimates.

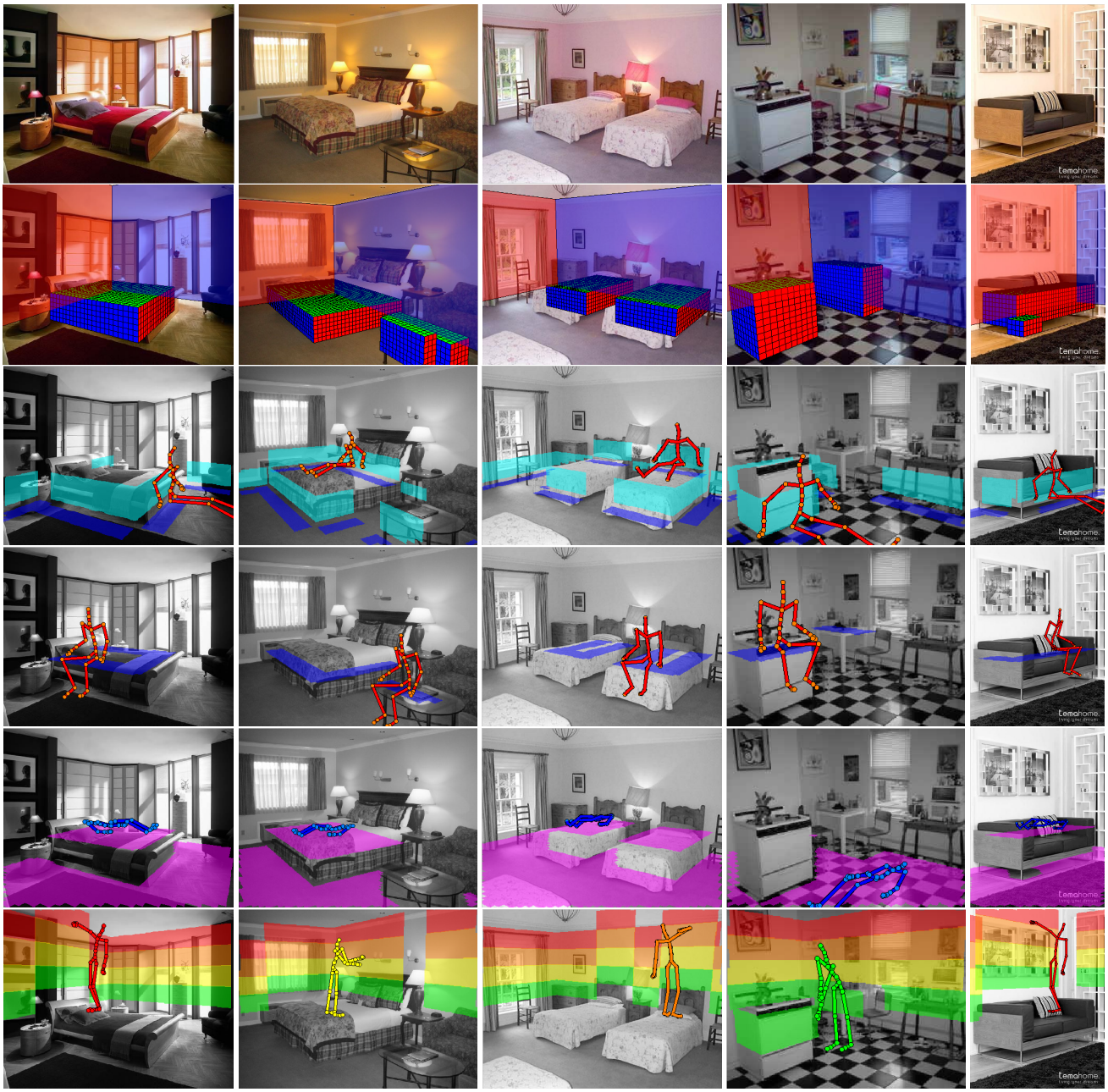


Figure 6: Qualitative performance of our approach on images with low calibration errors. The images in the first row are the input to our algorithm. The second row shows our estimated 3D scene geometry. The third row shows the possible pelvic joint and back support locations in blue and cyan respectively for the "sitting reclined" pose. The fourth row shows the possible pelvic joint locations in blue for the "sitting upright" pose. The fifth row shows the locations where a human's back can rest when "laying down." The last row shows the vertical surfaces a person's hand can touch from a standing position for the "reaching" pose, color coded to indicate the corresponding pose. Each scene also includes a representative stick figure for each pose.

ting. Specifically, we use the image features and multiple segmentations classifier of [13]. We use 50 training images for each classifier.

	Baseline	Our Approach
Reaching	0.3733	0.5431
Laying Down	0.4189	0.4786
Sitting Upright	0.0451	0.2081
Sitting Reclined	0.0056	0.1222

Table 1: Quantitative comparison of our approach with an appearance-based classifier.

Table 1 shows the performance of our approach compared to the baseline appearance classifier on each of the four classes. To evaluate the quality of a result, we use the pixel-wise overlap score metric. Our approach outperforms the appearance-based classifier in all categories. While the appearance-based classifier does a decent job in predicting valid locations for "laying down" and "reaching," it completely fails for both "sitting upright" and "sitting reclined." This is because predicting actions such as sitting require

Figure 8: Comparison of valid poses for adults and children.

global reasoning which appearance-based approaches fail to capture.

7.3. Subjective Affordances

People come in all shapes and sizes. This natural variation dictates how we interact with our environment. To illustrate how the same objects in a scene can afford different actions for different people, we conducted a proof-of-concept experiment using a 6ft tall adult and a 3ft tall child on a scene with annotated geometry.

Figure 8 shows valid "sitting reclined" and "reaching" poses for an adult and a child found automatically by our human-scene interaction algorithm. Note that a child would be capable of sitting with its legs up on the couch; however, for an adult to have the same pose, they would have to rest their legs on a second surface of support (the coffee table). Similarly, the child is able to stand on the narrow base of support (the back of the couch), in order to reach the same height the adult can by standing on the ground.

8. Using the Joint Human/Scene Space

The big question is: "What can we do, once we have a joint space of human pose and scene geometry?" We believe that our approach not only provides a fresh perspective on scene understanding that looks at predicting potential actions; but, it can be a vital link for solving several traditional vision problems. Three of these possible applications are:
1) Priors for Actions Recognition: Predicting the set of potential interactions and possible poses given an environment provides strong priors for action recognition.
2) Priors for Object Recognition: Valid pose positions could be used as a prior on the locations of objects in a scene. For example, one can compute the set of possible hand locations for all reaching poses, which provides a prior as to where manipulable objects are likely to be found.
3) Improving 3D Geometry Estimation: Indoor environments are designed to afford our daily activities. Knowledge of what tasks a human performs in an environment defines a set of poses which are known to be possible in that location. Thus, we can close the loop and use these poses to improve 3D geometry estimates.

Evaluating these possible applications in a comprehensive manner is beyond the scope of this paper. But, our preliminary experiments on pose prediction suggest that our approach could be useful for the vision tasks described above. We hope that this work opens the door for future exploration into how humans physically interact with their environment.

Acknowledgments: This research is supported by MURI Grant N000141010934. The authors would like to thank Varsha Hedau and David Lee for providing their results on the indoor scene dataset.

References

[1] CMU motion capture database. http://mocap.cs.cmu.edu. 1963
[2] Google 3D warehouse. http://sketchup.google.com/3dwarehouse. 1964
[3] Character motion systems. In *SIGGRAPH-Course 9*, 1994. 1962
[4] M. Brady, P. E. Agre, D. J. Braunegg, and J. H. Connell. The mechanics mate. *Advances in AI*, 1985. 1962
[5] J. Craig. *Intro. to Robotics: Mech. & Control*. MIT Press, 1989. 1961
[6] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 1961
[7] J. Gibson. The ecological approach to visual perception. *Boston: Houghton Mifflin*, 1979. 1962
[8] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *ECCV*, 2010. 1961, 1962
[9] A. Gupta, A. Kembhavi, and L. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *PAMI*, 2009. 1962
[10] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. In *Cambridge Press*, 2000. 1965
[11] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In *ICCV*, 2009. 1961, 1962, 1963, 1965, 1966
[12] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *ECCV*, 2010. 1962, 1963
[13] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. In *IJCV*, 2007. 1961, 1962, 1967
[14] K. Koffka. Principles of gestalt psychology. *NY*, 1935. 1962
[15] D. Lee, A. Gupta, M. Hebert, and T. Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *NIPS*, 2010. 1962, 1963, 1965, 1966
[16] D. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In *CVPR*, 2009. 1961, 1962, 1963
[17] U. Neisser. Direct recognition and perception as distinct perceptual systems. *Ann. Meet. of Cognitive Sci. Society*, 1989. 1962
[18] S. Palmer. Vision science: Photons to phenomenology. *MIT Press*, 1999. 1962
[19] E. Rivlin, S. J. Dickinson, and A. Rosenfeld. Recognition by functional parts. In *CVIU*, 1995. 1962
[20] A. Saxena, M. Sun, and A. Y. Ng. Make3D: Learning 3D scene structure from a single still image. *PAMI*, 2009. 1961
[21] L. Stark and K. Bowyer. Achieving generalized object recognition through reasoning about association of function to structure. In *PAMI*, 1991. 1962
[22] H. Wang, S. Gould, and D. Koller. Discriminative learning with latent variables for cluttered indoor scene understanding. In *ECCV*, 2010. 1961
[23] P. H. Winston, T. O. Binford, B. Katz, and M. Lowry. Learning physical description from functional definitions, examples, and precedents. *MIT Press*, 1984. 1962
[24] K. Yamane, J. J. Kuffner, and J. K. Hodgins. Synthesizing animations of human manipulation tasks. *ACM. Trans. on Graphics*, 2004. 1962
[25] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 1962
[26] S. Yu, H. Zhang, and J. Malik. Inferring spatial layout from a single image via depth-ordered grouping. In *CVPR Workshop*, 2008. 1961