Noname manuscript No.
(will be inserted by the editor)

People Watching

Human Actions as a Cue for Single View Geometry

David F. Fouhey · Vincent Delaitre · Abhinav Gupta · Alexei A. Efros · Ivan Laptev · Josef Sivic

Received: date / Accepted: date

Abstract We present an approach which exploits the coupling between human actions and scene geometry to use human pose as a cue for single-view 3D scene understanding. Our method builds upon recent advances in still-image pose estimation to extract functional and geometric constraints on the scene. These constraints are then used to improve single-view 3D scene understanding approaches. The proposed method is validated on monocular time-lapse sequences from YouTube and still images of indoor scenes gathered from the Internet. We demonstrate that observing people performing different actions can significantly improve estimates of 3D scene geometry.

Keywords Scene understanding · action recognition · 3D reconstruction

1 Introduction

The human body is a powerful and versatile visual communication device. For example, pantomime artists can convey elaborate storylines completely non-verbally and without props, simply with body language. Indeed, body pose, gestures, facial expressions, and eye movements are all known to communicate a wealth of information about a person, including physical and mental state, intentions, reactions, etc. But more than that, observing a person can inform us about the *surrounding environment* with which the person interacts.

Consider the two people detections depicted in Figure 1. Can you tell which one of the three scenes these detections came from? Most people can easily see that it is room A. Even though this is only a static image, the actions and poses of the disembodied figures reveal a lot about the geometric structure of the scene. The pose of the left figure reveals a horizontal surface right under its pelvis ending abruptly at its knees. The right figure's pose reveals a ground plane under its feet as well as a likely horizontal surface near the hand location. In both cases we observe a strong physical and functional coupling between people and the 3D geometry of the scene. In this work, we aim to exploit this coupling.

This paper proposes to use human pose as a cue for 3D scene understanding. Given a set of one or more images from a static camera, the idea is to treat each person as an "active sensor," or probe that interacts with the environment and in so doing carves out the 3D free-space in the scene. We reason about human poses following J.J. Gibson's notion of *affordances* [19] – each pose is associated with the local geometry that permits or *affords* it. This way, multiple poses in space and time can jointly discover the underlying 3D structure of the scene.

In practice, of course, implementing this simple and elegant scenario would be problematic. First of all, the underlying assumption that the humans densely explore the entire observed 3D scene is not realistic: in many scenes, humans may not interact with certain regions

D.F. Fouhey, A. Gupta, A.A. Efros
Carnegie Mellon University
Robotics Institute
5000 Forbes Avenue, Pittsburgh PA 15213
E-mail: dfouhey@cs.cmu.edu
A.A. Efros now with the EECS department at UC Berkeley

V. Delaitre, I. Laptev, J. Sivic
INRIA, WILLOW project,
Département d'Informatique de l'École Normale Supérieure,
ENS/INRIA/CNRS UMR 8548.
23, Avenue d'Italie
75013 Paris, France

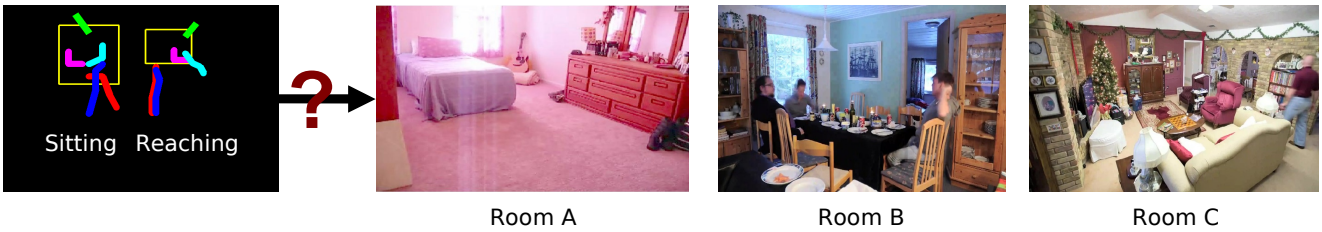


Fig. 1 What can human actions tell us about the 3D structure of a scene? Quite a lot, actually. Consider the people depicted on the left. They were detected in a time-lapse sequence in one of rooms A, B, or C. Which room did they come from? See the text for the answer.

for months. But more problematic is the need to recover high-quality 3D pose information for all people in an image. While several very promising 2D pose estimation approaches exist [1,32,53], and while it is possible to use anthropometric constraints to lift the poses into 3D [49], the accuracy of these methods is still too low to be used reliably.

As a result, we take a soft, hybrid approach that integrates appearance cues as well. We first employ the single-view indoor reconstruction method of Hedau *et al.* [27] which produces a number of possible 3D scene hypotheses. We then use existing human detection machinery to generate pose candidates. The crux of our algorithm is in simultaneously considering the appearance of the scene and perceived human actions in a robust way to produce the best 3D scene interpretation given all the available evidence. We evaluate our approach on both time-lapses and still images taken from the Internet, and demonstrate significant performance gains over state-of-the-art appearance-only methods. We additionally demonstrate the viability of using humans as a cue without appearance and provide substantial analysis of how and when observing humans helps us better understand scenes.

1.1 Background

Our goal is to understand images in terms of 3D geometry and space. Traditional approaches in computer vision for 3D understanding have focused on using correspondences and multiple view geometry [26]. While these methods have been successful, they are not applicable when only a single view of the scene is available; accordingly, they are incapable of understanding the great wealth of consumer and historic photographs not captured with multiple views or depth sensors. Reasoning about this sort of data is an enormous challenge for computers and an open research question since it is a wildly underconstrained problem. Nonetheless, it is trivial for any human being. Since humans can infer scene structure from a single image, single-view reconstruction is thus a necessary step towards vision systems with human-like capabilities. Furthermore, 3D scene estimates from a single image not only provide a richer interpretation of the image but also improve performance of traditional single-image tasks such as object detection [28,30].

Past work on single-image 3D understanding has overcome the underconstrained nature of the problem in a variety of ways. For instance, Kanade demonstrated the recovery of 3D shapes from a single image in [33] with the overarching constraint that "regularities observable in the picture are not by accident, but are some projection of real regularities," along with a variety of regularities such as parallelism and skewed symmetry. Advances in computing power as well as increased availability of data have led to work that learns to produce 3D understandings from data. This has led to methods that take a single image and recover the dominant 3D directions and horizon line of scenes [3], surface orientations [17,29,40], volumetric objects [24,47,52], or depth [34,45]. In many of these, constraints used in the past (e.g., parallelism) are integrated into the models' objective functions.

There has been particular success in the field of indoor single-view 3D understanding. This has been made possible by the fusion of effective learning techniques for single image 3D prediction such as [29] and low-dimensional (i.e., highly constrained) models for 3D layout specifically tailored for indoor images [5,8,9,27,37, 48,51,55]. The constraints involved usually concern the human-made nature of the scene: almost all approaches make the Manhattan-world assumption in which there are three orthogonal scene directions, and in most cases, it is assumed that the room can be roughly modeled as a cuboid. However, although they assume a human-centric scene structure, each of these approaches treats humans as clutter rather than as a cue: for all previous work, the best and arguably correct reaction to a human in the scene is to classify her as clutter and to ignore her pixels in room layout estimation. In fact, despite being fundamentally dependent on the human-centric nature of the scenes, all recent work on room layout predic-

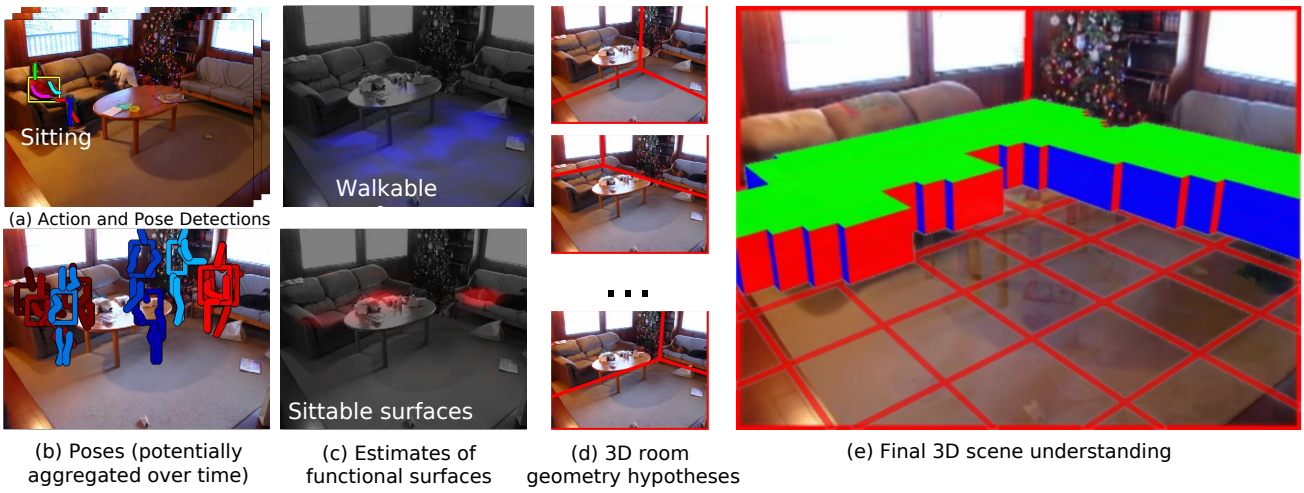


Fig. 2 Overview of the proposed approach. We propose the use of both appearance and human action cues for estimating single-view geometry. Given an input image or set of input images taken by a fixed camera, we estimate human poses in each image (a), yielding a set of human-scene interactions (b), which we aggregate over time (for time-lapses). We use these to to infer functional surfaces (c) in the scene, such as sittable (red) and walkable (blue). We simultaneously generate room hypotheses (d) from appearance cues alone. We then select a final room hypothesis and infer the occupied space in the 3D scene using both appearance and human action cues.

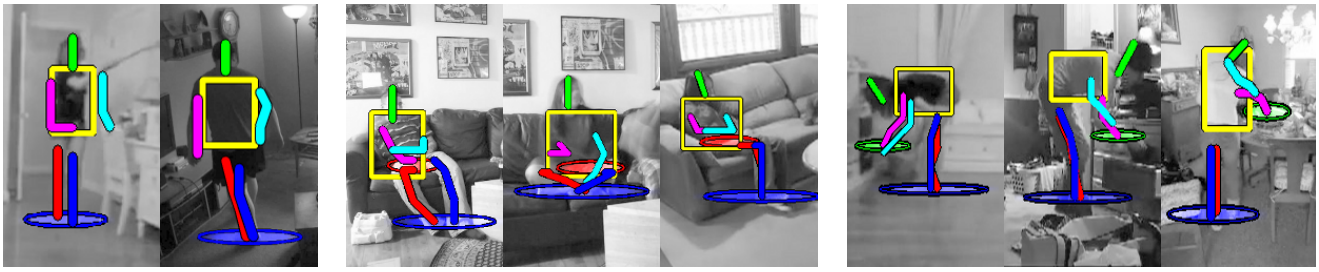
tion has been evaluated on datasets containing exactly zero people. Given that humans and their activities are often the primary motivation for documenting scenes and that human scenes are constructed for humans, this seems unnatural. This work aims to demonstrate that humans are not a nuisance, but rather another valuable source of constraints.

Other work on the interface between humans and image understanding has mostly focused on modeling these constraints at a semantic level [11,23,50]. For example, drinking and cups are functionally related and therefore joint recognition of the two should improve performance. Semantic-level constraints have been also shown to improve object discovery and recognition [18, 41,50], action recognition [11,12,23,35], and pose estimation [22,54]. Recently Delaitre *et al.* [10] proposed the use of poses for semantic segmentation of scenes; like ours, their work also uses poses as a cue, but it solves the complementary problem of giving each pixel in an image a semantic label (e.g., chair), not improving estimates of 3D scene geometry.

In this paper we specifically focus on modeling relationships at a physical level between humans and 3D scene geometry. In this domain, most earlier work has focused on using geometry to infer human-centric information [20,25], or the question "what can a human do with a given 3D model". For instance, Gupta *et al.* [25] argued that functional questions such as "Where can I sit?" are more important than categorizing objects by name, and used estimated 3D geometry in images to infer Gibsonian affordances [19], or "opportunities for interaction" with the environment. Jiang *et al.* [31] used the sizes and poses of humans to infer human-object affordances from 3D scenes containing no humans.

Our work focuses on the inverse of the problem addressed in [20,25]: we want to observe human actors, infer their poses and then use the functional constraints from these poses to improve 3D scene understanding. Our goal is to harness the recent advances in person detection and pose estimation [1,4,14,32,53], and design a method to improve single-view indoor geometry estimation. Even though the building blocks of this work, human pose estimation and 3D image understanding, are by no means perfect, we show that they can be robustly combined. We also emphasize our choice of the monocular case, which sets our work apart from earlier work on geometric reasoning using human silhouettes [21] in multi-view setups. In single-view scenarios, the focus has been on coarse constraints from person tracks [36, 44,46], whereas we focus on fine-grained physical and functional constraints using human actions and poses.

A preliminary version of this work appeared as [16]. We clarify many technical details omitted in the previous version, present results on substantially extended datasets, and offer an in-depth analysis of how, why, and when observing humans can improve understanding of the 3D geometry of scenes.



(a) Standing (b) Sitting (c) Reaching

Fig. 3 Example action detection and pose estimation results with the articulated model. The predicted surface contact points are shown by ellipses: blue (walkable), red (sittable), green (reachable). Shown actions are: standing (1-2), sitting (3-5), and reaching (6-8).

2 Overview

Our work is an attempt to marry human action recognition with 3D scene understanding. We have made a number of simplifying assumptions. We limit our focus to indoor scenes: they allow for interesting human-scene interactions and several successful approaches exist specifically for estimating their geometry from a single view [27,37]. We use a set of commonly observed physical actions, reaching, sitting, and walking, to provide constraints on the free and occupied 3D space in the scene. To achieve this, we manually define surface constraints provided by each action, e.g., there should be a sittable horizontal surface at the knee height for the sitting action. We adopt a geometric representation that is consistent with recent methods for indoor scene layout estimation. Specifically, each scene is modeled in terms of the layout of the room (walls, floor, and ceiling) and the 3D layout of the objects. It is assumed that there are three principal directions in the 3D scene (the Manhattan world assumption [6]) and therefore estimating a room involves fitting a parametric 3D box that is aligned with the vanishing points.

While temporal information can be useful for detecting human actions and imposing functional and geometrical constraints, in this work, we only deal with still images and time-lapse videos with no temporal continuity. Time-lapses are image sequences recorded at a low framerate, e.g., one frame a second. Such sequences are often shot with a static camera and show a variety of interactions with the scene while keeping the static scene elements fixed. People use time lapses to record and share summaries of events such as home parties or family gatherings. Videos can nonetheless be used in the proposed framework without any modifications by ignoring the temporal information and treating them as a time-lapse, or by substituting our single-frame pose estimators with approaches that integrate temporal information [2,39,42].

Time-lapse data is ideal for our experiments for many reasons. In our case, the time discontinuity works in our favor as it naturally compresses highly diverse person-scene interactions into a small number of frames: a time-lapse video lasting a few minutes may show many hours of events. Moreover, the static nature of the underlying scene enables joint reasoning about multiple images without solving for camera pose to find a common reference frame. This lets us focus on the core of our problem, rather than a structure-from-motion preprocessing step. Finally, the time-lapses also enable us to test our method on realistic data with non-staged activities in a variety of natural environments: time-lapses are captured by consumers in their daily living environments, frequently with viewing angles similar to consumer still photographs. Our datasets, for instance, contain time-lapses gathered from a consumer video sharing site, YouTube.com.

An overview of our approach is shown in Figure 2. First, we detect humans performing different actions in the image and use the inferred body poses to extract functional regions in the image such as sittable and reachable surfaces (Section 3). For time-lapses, we accumulate these detections over time for increased robustness. We then use these functional surface estimates to derive geometrical constraints on the scene. These constraints are combined with an existing indoor scene understanding method [27] to predict the global 3D geometry of the room by selecting the best hypothesis from a set of hypotheses (Section 4.1). Once we have the global 3D geometry, we can use these human poses to reason about the free-space of the scene (Section 4.2).

3 Local Scene Constraints from People's Actions

Our goal is to predict functional image regions corresponding to *walkable*, *sittable* and *reachable* surfaces by analyzing human actions in the scene. We achieve this

by detecting and localizing people performing the three different actions (standing, sitting, reaching) and then using their pose to predict *contact points* with the surfaces in the scene. For time-lapses, contact points are aggregated over multiple frames to provide improved evidence for the functional image regions. We illustrate these contact points on detected humans in Fig. 3.

Given a person detected performing an action, we predict contacts with surfaces as follows: (i) for **walkable** surfaces we define a contact point as the mean location of the feet position, and use all three types of actions; (ii) for **sittable** surfaces, we define a contact point at the mean location of the hip joints, and consider only sitting actions; and (iii) for **reachable** surfaces, we define a contact point as the location of the hand further from the torso, and use only reaching actions. These surfaces are not mutually exclusive (e.g., the tops of beds are sittable and reachable) and are estimated independently. To increase robustness to mistakes in localization, we place a Gaussian at the contact points of each detection and weight the contribution of the pose by the classifier confidence; one can also equivalently view this as each contact point voting for the properties of the scene with higher weight placed on nearby locations and on detections with high confidence. The standard deviation of each Gaussian is set to a fraction of the detection bounding box, 1/4 in X- and 1/40 in Y-direction, respectively; we use the bounding box dimensions to automatically scale our region of uncertainty with human proportions. This yields probability maps h for the different types of functional image regions, as illustrated in Figures 2c and 5c,d.

Since a sitting detector may also respond to standing people, we discriminate between different actions by converting the detection score of each model into a probability by fitting a decreasing exponential law on their firing rate. Action classification is performed with non-maximum suppression: if bounding boxes of several detections overlap irrespective of action class, the detection with the highest calibrated response is kept.

Our approach is agnostic to the particular method of pose detection and only requires a detector that produces a class (e.g., sitting) as well as estimates for the relevant joints of the human (e.g., pelvic joint, feet). In this work, we use two complementary approaches. We build primarily on the articulated pose model of Yang and Ramanan [53]. Here, we employ the model for detecting human action by training a separate model for each of the three actions. To supplement the articulated model, we use the deformable parts model (DPM) of Felzenszwalb *et al.* [14] for sitting and standing: the low variance of the relevant joints of these actions (e.g., feet for standing) enable us to accurately approximate

Fig. 4 Example detections with the deformable part models (top row: sitting; bottom row: standing) and approximated joint locations from bounding boxes (red: pelvic joint/sittable; blue: feet/walkable)

poses by simply transferring a fixed pose with respect to the bounding box. We find that these two methods have complementary strengths and error modes. Examples of detected actions together with estimated body pose configurations and predicted contact points for the articulated model are shown in Figure 3 and for the deformable parts model in Figure 4.

The articulated pose estimator and deformable parts model are trained and used separately and produce independent estimates of functional regions. These functional regions are integrated separately in our room layout ranking function in Equation 2.

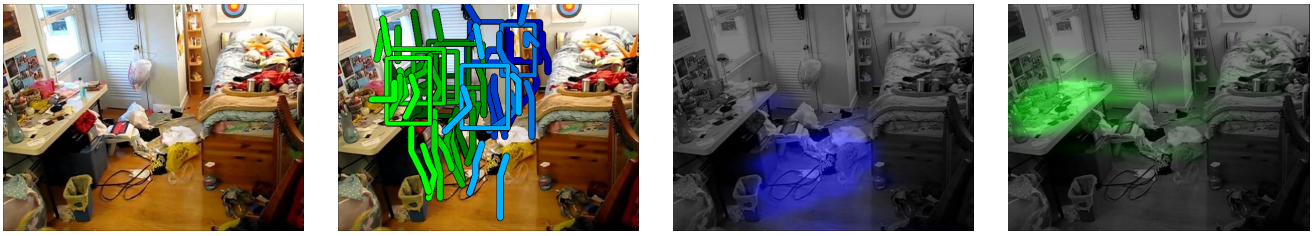
4 Space Carving Via Humans

In the previous section we discussed how we estimate human poses and functional regions such as sittable and walkable surfaces. Using the inferred human poses, we now ask: "What 3D scene geometry is consistent with these human poses and functional regions?" We build upon [25], and propose three constraints that human poses impose on 3D scene geometry:

Containment. The volume occupied by a human should be inside the room.

Free space. The volume occupied by a human cannot intersect any objects in the room. For example, for a standing pose, this constraint would mean that no voxels below 5ft can be occupied at standing locations.

Support. There must be object surfaces in the scene which provide sufficient support so that the pose is physically stable. For example, for a sitting pose, there must exist a horizontal surface beneath the pelvis (such as a chair). This constraint can also be written in terms of the functional regions; for example, the backpro-



(a) Input Image (b) Pose Detections (c) Walkable (d) Reachable

Fig. 5 Predicting functional image regions. (a) An image from a time-lapse sequence. (b) Overlaid example person detections from different frames: standing (blue), reaching (green). (c,d) Probability maps of predicted locations for (c) walkable and (d) reachable surfaces. Note that the two functional surfaces overlap on the floor.

jected sittable regions must be supported by occupied voxels in the scene.

Our goal is to use these constraints from observed human poses to estimate room geometry and the occupied voxels in the scene. Estimating voxels occupied by the objects in the scene depends on the global 3D room layout as well as the free-space and support constraints. On the other hand, estimating 3D room layout is only dependent on the containment constraint and is independent of the free-space and support constraints. Therefore, we use a two-step process: in the first step, we estimate the global 3D room layout, represented by a 3D box, using appearance cues and the containment constraints from human actors. This is done by ranking a large collection of room hypotheses and selecting the top-ranked hypothesis. In the second step, we use the estimated box-layout to estimate the occupied voxels in the scene. Here, we combine cues from scene appearance and human actors to carve out the 3D space of the scene.

4.1 Estimating Room Layout

Given an image and the set of observed human poses, we want to infer the global 3D geometry of the room. We build on the approach of Hedau *et al.* [27], which ranks a collection of vanishing-point aligned room hypotheses according to their agreement with appearance features. However, estimating the room layout from a single view is a difficult problem and it is often almost impossible to select the right layout using appearance cues alone: frequently, there are a handful of top hypotheses with inadequate evidence to decide which is correct. We propose to further constrain the inference problem by using the containment constraint from human poses. This is achieved with a scoring function that uses appearance terms as in [27] as well as terms to evaluate to what degree the hypothesized room layout is coherent with observed human actors.

Given input image features x and the observed human actors H (represented by functional surface probability maps $h \in H$), our goal is to find the best room layout hypothesis $y^* \in \mathcal{Y}$. We use the following scoring function to evaluate the coherence of image features and human poses with the hypothesized room layout y:

$$f(x, H, y) = \alpha_\psi \psi(x, y) + \alpha_\phi \phi(H, y) + \alpha_\rho \rho(y), \quad (1)$$

where $\psi(x, y)$ measures the compatibility of the room layout configuration y with the estimated surface geometry computed using image appearance, $\phi(H, y)$ measures compatibility of human poses and room layout, and $\rho(y)$ is a regularizing penalty term on the relative floor area that encourages smaller rooms; the αs trade-off between the compatability terms.

As we build upon the code of Hedau *et al.*, the first term, $\psi(x, y)$ is the scoring function learned via Eqns. 3-4 of [27]. The function uses global appearance cues, such as detected straight lines and the per-pixel classifier predictions for different surface labels (walls, floor, ceiling). The second term enforces the containment constraints and expands as

$$\phi(H, y) = \sum_{h \in H} \alpha_{\phi,h} \varphi(\zeta(h), y), \quad (2)$$

where $\zeta(h)$ is the mapping of functional surfaces onto the ground plane and φ measures the normalized overlap between the projection and floor in the hypothesized room layout. Intuitively, $\phi(H, y)$ enforces that the projection of both the human body and the objects it is interacting with should lie inside the room. The $\alpha_{\phi,h}$ terms trade off the weightings of the functional surfaces. In the current system, we do not enforce that the body does not intersect the ceiling, although this is not an issue in practice. We approximate $\zeta(h)$ by using the feet locations of detected actors, which produces accurate results for our action vocabulary. Finally, the term $\rho(y) = -\max(0, (A - M)/M)$ imposes a penalty for excessive floor area A, measured with respect to the

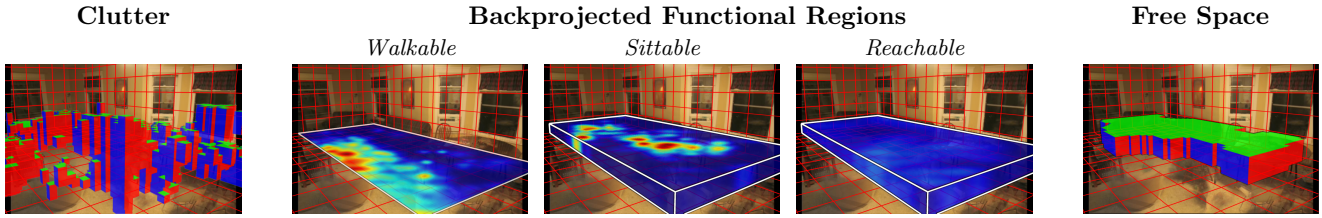


Fig. 6 We estimate the free space of the scene by taking the backprojected clutter map and refining it with backprojected functional regions. Voxels above all backprojected functional regions recieve votes against being occupied due to the free space constraint; voxels below reachable and sittable regions (i.e., within the volumes) receive votes in favor of being occupied due to the support constraint. The result of combining all cues is shown on the right.

minimum floor area M out of the top three hypotheses. We include this regularization term since $\phi(H, y)$ can only expand the room to satisfy the containment constraint.

The relative weights α on the terms are learned in a leave-one-out fashion. One term, α_ψ, is held constant, and grid search is performed on a fairly coarse grid (2^i for $i = \{-4, \ldots, 2\}$ for the other αs. For the regularization term, we include the additional option of setting α_ρ to 0. We chose the weighting that results in the highest mean performance when choosing the top-ranked room on the training set. However, we note that the system is fairly insensitive to the particular αs: on no parameter setting does the system produce worse results than the appearance-only system on any dataset.

We select our room interpretation from the same hypothesis pool $\mathcal{Y}$ as [27]. First, a modified version of the vanishing point detector of Rother [43] estimates three orthogonal vanishing points in the scene. The discrete pool of hypotheses is generated by discretizing the space of all vanishing-point aligned boxes. We rank each hypothesis with Equation 1 and return the top-scoring hypothesis as our predicted layout.

4.2 Estimating Free Space in the Scene

Once we have estimated the room layout we now estimate the voxels occupied by objects. However, this is a difficult and ill-posed problem. Hedau *et al.* [27] use an appearance based classifier to estimate pixels corresponding to objects in the scene. These pixels are then back-projected under the constraint that every occupied voxel must be supported. Lee *et al.* [37] and Gupta *et al.* [25] further constrain the problem with domain-specific cuboid object models and constraints such as "attachment to walls." We impose functional constraints: a human actor carves out the free space and support surfaces by interacting with the scene.

The room layout and camera calibration gives a cuboidal 3D voxel map in which we estimate the free space. We first back project the clutter mask of Hedau *et al.* [27], and then incorporate constraints from different human poses to further refine this occupied voxel map. Specifically, we backproject each functional region h at its 3D height, yielding a horizontal slice inside the voxel map. Because our classes are fine-grained, we can use human dimensions (waist height) for the heights of the sitting and reaching surfaces. This slice is then used to cast votes above and below in voxel-space: votes in favor of occupancy are cast in the voxels below; votes against occupancy are cast in the voxels above. This is illustrated in Fig. 6. The final score for occupancy of a particular voxel is a linear sum of these votes; as the result is probabilistic, to produce a binary interpretation as shown in the figures, we must threshold the results.

5 Experiments

In this section, we describe experiments done to validate our contributions. We introduce the experimental setup, the datasets, and raw quantitative results as well as some qualitative results. A detailed analysis of how observing people changes room layout estimation is presented in Section 6.

As preface, we note that the primary contribution of this work is the demonstration that humans can serve as a valuable cue for single-view geometry problems in practice. Although we present a way of integrating functional surfaces into single-view reasoning with a particular appearance-based system operating in a particular paradigm, there are many other ways of recovering the layout, surfaces, or 3D structure of a room from a single view. The purpose of this work is not to demonstrate that our particular system out-performs all other approaches; instead, it is to demonstrate that affordance-based cues offer complementary evidence to appearance-based ones. Accordingly, we design our experiments and analysis to investigate how our system performs relative to a system using the same appearance features and set of hypotheses.

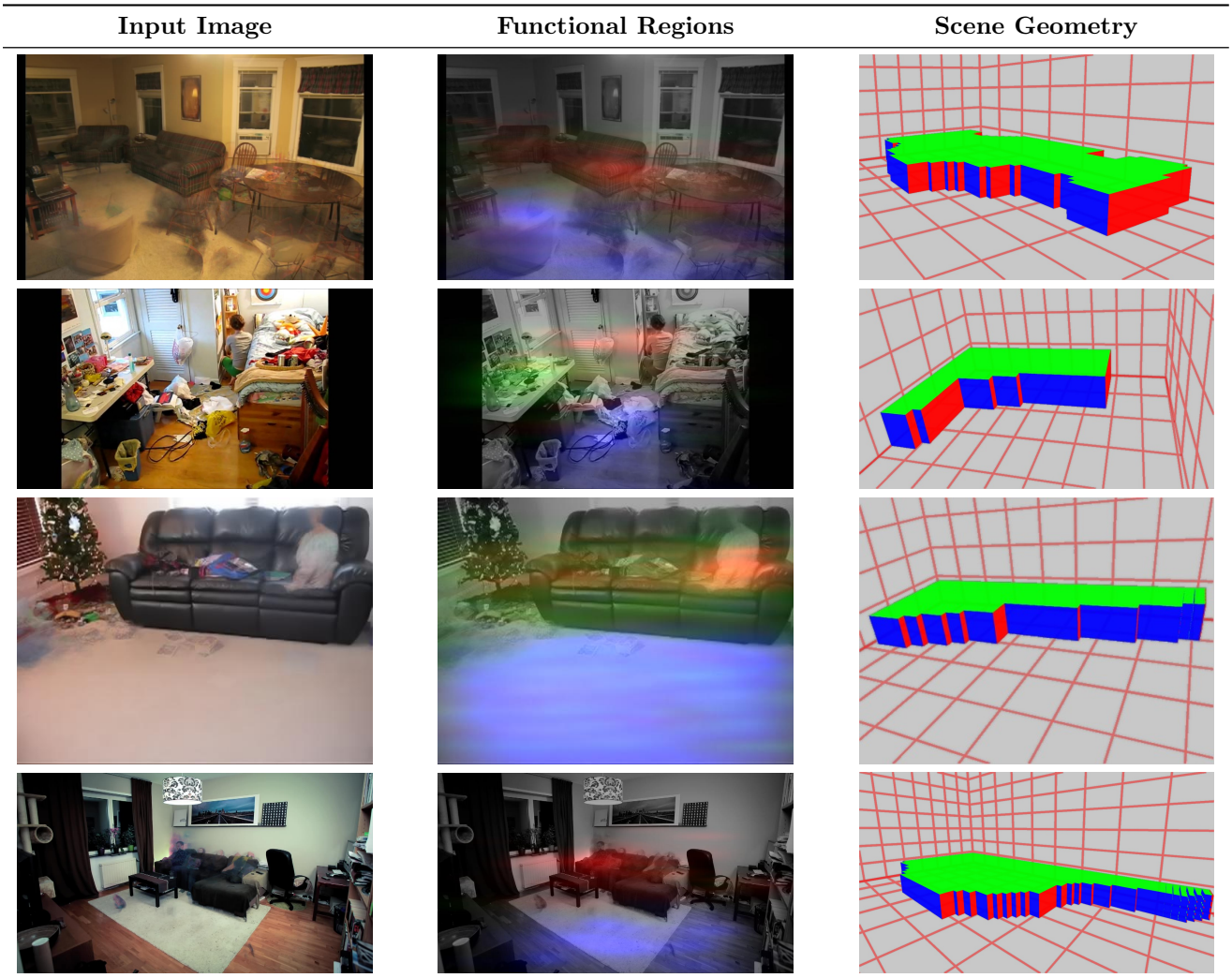
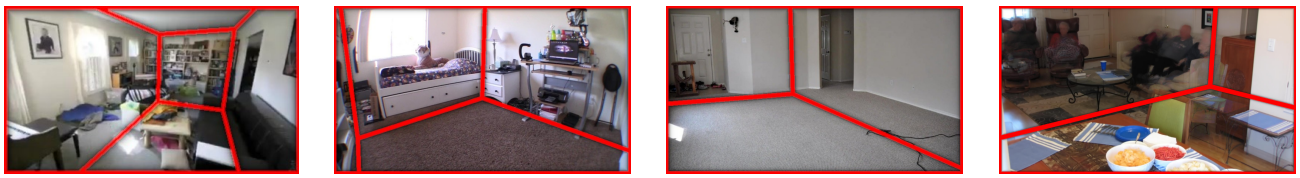


Fig. 7 Example time-lapse sequence results: given an input image, we use functional regions (walkable: blue; sittable: red; reachable: green) to constrain the room layout; having selected a layout, we can also infer a more fine-grained geometry of the room via functional reasoning.



(a) Appearances Only (Hedau *et al*).

(b) Appearances + People (Our approach).

Fig. 8 Timelapse experiment: A comparison of (a) the appearance only baseline [27] with (b) our improved room layout estimates. In many cases, the baseline system selects small rooms due to large amounts of clutter. On the right, even though the room is not precisely a cuboid, our approach is able to produce a better interpretation of the scene.

Baselines and evaluation criteria. For both time-lapses and single images, we compare our estimates of room geometry to a number of baselines. Our primary baseline is the appearance-only system of Hedau *et al.* [27], which we use to provide appearance cues. To provide context, we also include another baseline, in which we impose the box model on the output of Lee *et al.* [38] that maximizes the agreement. Finally, to show that all methods are operating better than chance, we use location alone to predict the pixel labels, akin to a per-pixel prior: after resizing all scenes to common dimensions, we use the majority label in the training images for each pixel.

To quantitatively evaluate how well we estimate the layout of an image, we use the standard metric of per-pixel accuracy (i.e., treating the problem as a semantic segmentation one). Note that since the camera is fixed in time-lapses, the scene can be summarized with a single image (the one provided to the appearance-only approaches) and thus only a single annotation is needed. In some time-lapses, the camera is adjusted or zoomed slightly; this only impacts our approach, and not the apperance-only approaches. When aggregating a statistic over a dataset, we quantify our uncertainty regarding the statistic via boostrapped confidence intervals, which we compute with the Bias-Corrected Accelerated method [13], using 10,000 replicates.

Implementation details. We train detectors using the Yang and Ramanan model for all three actions [53] and the Felzenszwalb *et al.* model for sitting and standing. For the standing action, we use a subset of 196 images from [53] containing standing people. For sitting and reaching, we collect and annotate 113 and 77 new images, respectively. All images are also flipped, doubling the training data size. As negative training data we use the INRIA outdoor scenes [7], indoor scenes from [27], and a subset of Pascal 2008 classification training data. None of the three negative image sets contains people. On testing sequences, adaptive background subtraction is used to find foreground regions in each frame, which are used as the appearance for the time-lapse and to remove false-positive detections on the background. We also use geometric filtering similar to [30] to remove detections that significantly violate the assumption of a single ground plane.

5.1 Datasets

We test the proposed approach on consumer time-lapse videos and a collection of indoor still images. The data for both originates from the Internet and depicts challenging, cluttered, and non-staged scenes capturing one or more people engaged in everyday activities interacting with the scene. Comparison on existing datasets is impossible since previous work on room layout estimation has ignored people as a cue and has been evaluated on datasets composed entirely of unoccupied rooms.

Time-lapse data. For time-lapse videos, we present results on the dataset introduced in [16], as well as on the larger semantic segmentation dataset presented in [10], which contains the dataset of [16] as a subset. We refer to the dataset of [16] as the *People Watching* time-lapses, and that of [10] as the *Scene Semantics* time-lapses. We have re-labeled the *Scene Semantics* dataset for room layout prediction.

Both datasets were collected from Youtube using keywords such as "time-lapse," "living room," "party," or "cleaning." The *People Watching* dataset contains 40 videos with about 140,000 frames. The *Scene Semantics* dataset contains 146 videos, totaling about 400,000 frames.

Unlike the *People Watching* dataset, which contains largely unambiguous and cuboidal bedrooms and living rooms, the *Scene Semantics* dataset contains a wider variety of scene classes as well as many scenes that violate the assumptions of our method. These include straight-forward violations of explicit assumptions, such as that no more than three walls are visible. These also include violations of more subtle implicit assumptions, such as that the full body of a person will be within the frame and that the floor will occupy some reasonable fraction of the scene: 10% of scenes in the *Scene Semantics* dataset have floor coverage of 15% or less, as compared to none in the *People Watching* dataset. This leads to truncated detections on the bottom of the image that do not actually rest on the floor. In the handful of cases where the cuboidal model is explicitly violated, only the unambiguous parts are annotated: for instance if four walls are visible due to a fish-eye lens, only the floor and ceiling are annotated.

Still image data. Our previous work [16] introduced a dataset of 100 still images of indoor scenes with people. We expand this dataset to 500 images. These images were retrieved from the Internet with queries such as "living room" and "waiting room" with the criterion that they are roughly cuboidal and that they contain at least one person. These non-staged images depict celebrities, political figurs, and ordinary people engaged in everyday tasks, ranging from simply sitting and talking to having meetings or parties.

Table 1 Time-lapse Experiment: Average pixel accuracy for geometry estimation on the *People Watching* [16] and *Scene Semantics* [10] Time-lapse datasets. Our method achieves significant gains and using humans alone produces competitive performance.

	Location	**Appearance Only**		**People Only**	**Appearance + People**
		Lee *et al.*	Hedau *et al.*		
People Watching [16]	64.1%	70.4 %	74.9%	70.8%	**82.5%**
Scene Semantics [10]	61.4%	68.3 %	74.3%	70.7%	**77.4%**

5.2 Experiment 1: Time-lapse data

Figure 7 shows the performance of our approach on a set of time-lapses. The second column shows the probabilistic estimates of "walkable", "sittable" and "reachable" surfaces in blue, red and green respectively. We use these functional region estimates to select the best room hypothesis and estimate the free space of the scene, which is shown in the third column. These results show that human actors provide lot of information about the scene as they interact with it. For example, in the second row, even though the scene is cluttered, human reaching actions help us to infer a horizontal surface on the left. We also qualitatively compare the 3D room layout estimated by our approach to that of Hedau *et al.* [27]. Figure 8 shows some examples of the relative performance.

We present a summary of the quantitative results on both time-lapse datasets in Table 1. In both cases, our method is able to consistently improve on the baseline. On the *People Watching* dataset, our method averages a gain of 7.6% (bootstrapped 95% confidence interval: 4.5% to 11.3%) over the Appearance-alone baseline. Further, our performance is as good (within 5%) or better than the baseline in 92.5% of images. On the *Scene Semantics* dataset, our method averages a gain of 3.1% (bootstrapped 95% confidence interval: 1.1% to 5.1%) over the Appearance-only baseline. Further, on an individual image basis, our performance is generally about the same or better: in the overwhelming majority of cases, 88.6%, our system produces as good or better results.

To demonstrate the power of cues from people, we show results using human action cues alone to select room hypotheses. Specifically, we use only our human action compatibility and room size terms, ϕ and ρ, to rank the hypotheses. Since we have no appearance term to counter-balance the expansion preferred by our containment constraint, we modify ρ to use the minimum floor area from a larger number of top hypotheses. Specifically, this fraction is the top $n\%$ for $n = \{1, 5, 10, 25, 50\}$ selected via leave-one-out cross-validation. This can be thought of as aiming to select a small room among the top $n\%$ hypotheses that explain the people observed in the scene. Even with only people as cues, our system performs only about 4% worse on average than Hedau *et al.* and equivalently to Lee *et al.* on both time-lapse datasets.

Finally, we evaluate the use of functional regions as a cue for free space estimation on the *Scene Semantics* dataset. Since high-quality labels already exist for semantic segmentation from [10], we evaluate in the image plane, predicting whether a region is free or not on the floor. We linearly combine the functional regions with the clutter map as in Section 4.2 with the weighting learned by logistic regression. We characterize performance via average precision (AP) obtained by sweeping a threshold over classifier output. In comparison to the clutter map produced by [27], our approach obtains a boost of 6.7% AP (boostrapped 95% confidence interval: 4.6% to 8.9%) when combined with the clutter map. Further, using people alone as an indicator of free-space produces about the same accuracy as the predicted clutter map, a -0.5% comparative loss in AP (boostrapped 95% confidence interval -4.7% to 2.9%).

5.3 Experiment 2: Understanding single images

In the second experiment, we run the system on still images to see whether functional reasoning can improve room layout estimation with a single image. We report results on our new extended dataset of 500 images. We additionally report results on the dataset introduced in [16] for completeness. Our results show that functional constraints from human actions provide strong evidence of 3D geometry even in single images. Figure 9 shows a few examples of our estimated room geometry compared to Hedau *et al.* [27]. We are able to obtain substantial improvements over the appearance-only approach; in most cases in Figure 9, this gain comes when people disambiguate a cluttered scene where estimating the true extent of the room is difficult. Figure 10 shows examples of estimated 3D room geometry and the 3D occupied voxels.

We report quantitative results in Table 2. On the extended dataset, we we obtain a 1.26% improvement (bootstrapped 95% confidence interval: $0.61\% - 1.97\%$) and our performance is as good or better in 86.3% of

Table 2 Single Image Experiment: Average pixel accuracy for geometry estimation on single images. With even a single pose, our method achieves significant gains.

	Location	Appearance Only		Appearance + People	
		Lee *et al.*	Hedau *et al.*	Ours	with Ground Truth Poses
People Watching Single Images [16]	66.4%	71.3%	77.0%	79.6%	**80.8%**
Extended Single Images	62.5%	70.7%	76.0%	77.8%	**78.8%**

Fig. 9 Single Image Experiment: The correct person in the correct place can easily disambiguate complex scene interpretation problems. In the last example, although the vanishing points are inaccurate, we produce a more accurate intepretation.

cases. On average, our gain over the appearance-only baseline on single images is lower than in the time-lapse case although we obtain many scenes with substantial gains. This seeming drop in performance happens because a single frame gives us fewer chances to use functional reasoning to adjust our scene interpretation, resulting in equivalent interpretations with and without people (and even, as Table 2 shows, with the ground-truth poses). In subsequent analysis we more thorougly explore the nature of our gains in performance.

Fig. 10 Single Image Experiment: Functional reasoning to detect sittable surfaces in still images.

6 Discussion

We now present an analysis of how observing humans improves room layout prediction. We first show how observing humans changes layout estimation beyond summary statistics. We then analyze when our system succeeds and show how our performance varies with respect to the number of humans present and detected. Finally, we illustrate when our system does not succeed, and analyze a number of failure modes.

6.1 How do humans change layout estimation?

On average our system provides consistent performance gains over the appearance-only baseline; however, considering just mean performance hides important aspects of how performance varies. In particular, one could wonder if the performance gains are consistent or the result of a process with higher mean but high variance as well. The former is more reassuring than the latter since in the latter case whether performance improves is largely a product of chance; further, adding people as a cue should not hurt performance.

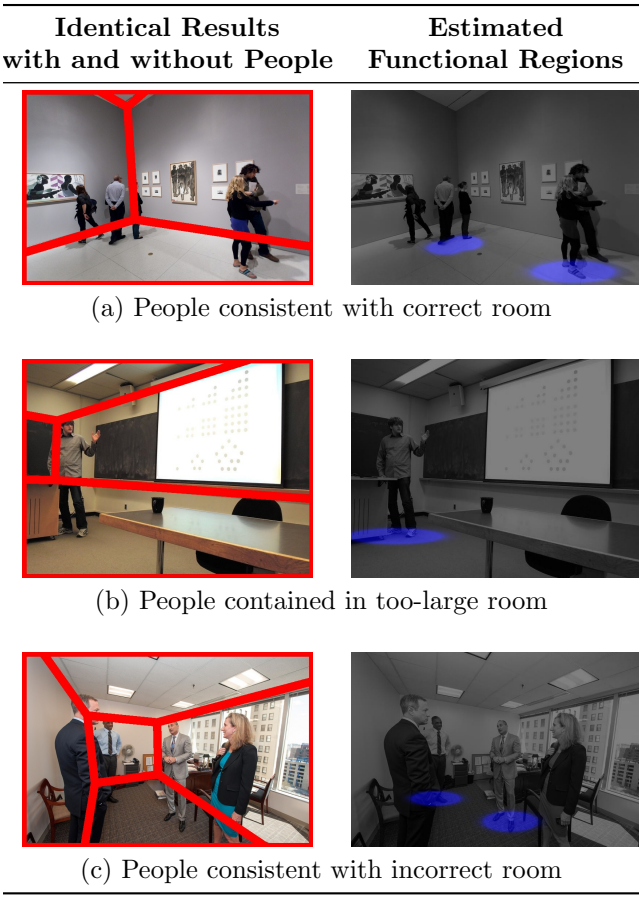
The peformance of the system is bi-modal: in one mode, the Appearance + People system produces similar results to the Appearance only system; in the other, functional reasoning produces different and systematically better results. We present an analysis of these two populations on the expanded still image dataset and

Table 3 Analysis of the scenes in which we produce similar (< 10% difference) and different predictions to the appearance only baseline. Across all datasets, we produce most of our gains from scenes in which appearance is inadequate.

	Still Images		**People Watching Time-lapses [16]**		**Scene Semantics Time-lapses [10]**	
	Different	Similar	Different	Similar	Different	Similar
Proportion	40.2%	59.8%	62.5%	37.5%	55.3%	44.7%
Appearance Acc.	71.6%	79.1%	71.9%	79.7%	71.4%	77.6%
Relative Acc.	3.0%	0.8%	11.5%	0.6%	5.3%	0.4%

both time-lapse datasets in Table 3. We define similar predictions as scenes in which the pixel-map output of the two systems differs by 10% or less, and different predictions as the rest. We present the proportion of the cases (i.e., how often they differ), the accuracy of the appearance-only system on each set of scenes, and the gain of the proposed system over the appearance-only baseline. Table 3 shows that our system consistently obtains most of its gains from the scenes in which appearance by itself is inadequate for reasoning. Compare the appearance-alone accuracies between the similar and different columns, as well as the relative performance: across all datasets, the scenes in which people help improve rooms are the ones in which appearance produces worse results.

This has important implications for the nature of our performance gains. It shows that the mean statistics presented in Tables 1 and 2 are not a large number of gains outweighing a large number of substantial losses leading to an improvement, but instead are systematic gains diluted down by largely unchanged (but correctly so) scenes. This no-difference mode stems from the nature of humans as a cue: unlike a new feature or learning algorithm, direct observation of humans can only act as a cue where humans are present in the scene. If the room is largely correct or the people in the scene are entirely contained within an incorrect interpretation, then functional reasoning cannot help. These cases are illustrated qualitatively in Figure 11.



(a) People consistent with correct room

(b) People contained in too-large room

(c) People consistent with incorrect room

Fig. 11 Illustrations of the cases where observing people produces identical results to using appearance cues alone. Top-to-bottom: frequently, the room is correct and our constraints do not change the ranking of hypotheses; sometimes the room's size is estimated as incorrectly large, and the containment constraint is satisfied; sometimes the room's size is too small, but all people are contained within the incorrect estimate. For clarity, the variance of the Gaussians for functional surface computation have been increased and the map has been rendered more strongly.

6.2 How many people do we need to see to succeed?

We now provide an analysis of how the number of people in a scene affects the performance of our system in terms of room layout estimation: does our system need to see a large number of people densely scattered through the scene in order to function or can our system achieve substantial gains with relatively few people in the scene? The latter is preferable to the former since our system would have little chance of working on most real world images if it required dozens of people spread throughout the scene. We answer the latter: our system does not rely on a dense crowd of people to find a room, but can use as little as a one peson to dramatically change a scene's interpretation. In the process, we analyze the connection between the number of people in the scene and the performance of our system relative to the appearance only baseline.

We show scatterplots of the relative performance of our system on the extended still image dataset against both the number of people present in the scene and

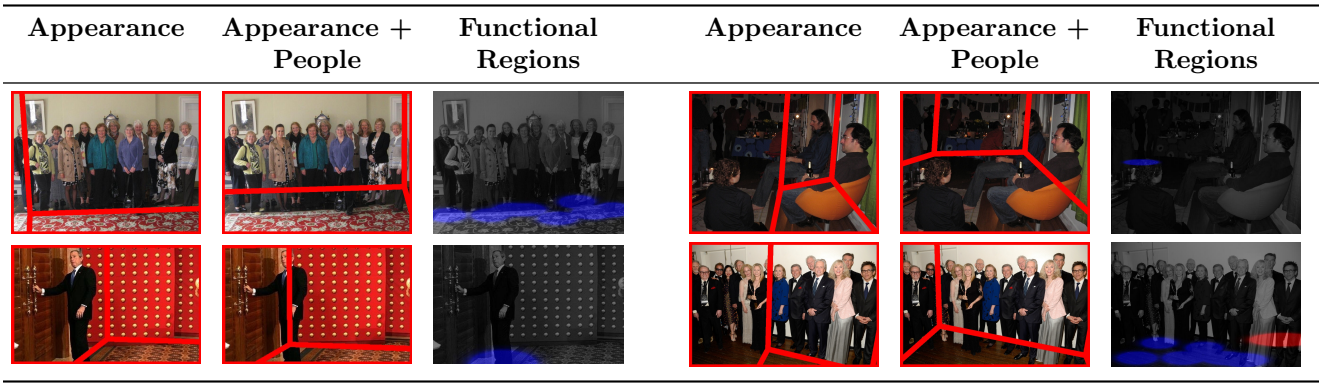
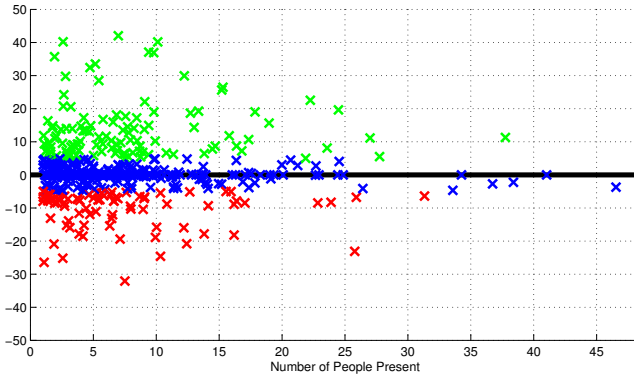
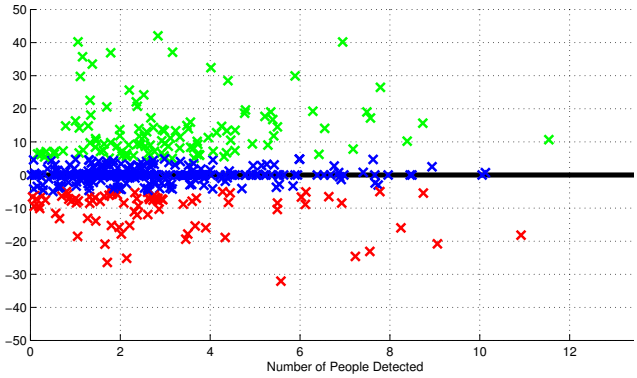


Fig. 13 Examples of small refinements (left) and large improvements (right) in scenes with small or large numbers of people detected. For clarity, the variances of the Gaussians for functional surface computation have been increased and the map has been rendered more strongly.



(a) Performance gain v. number of people present

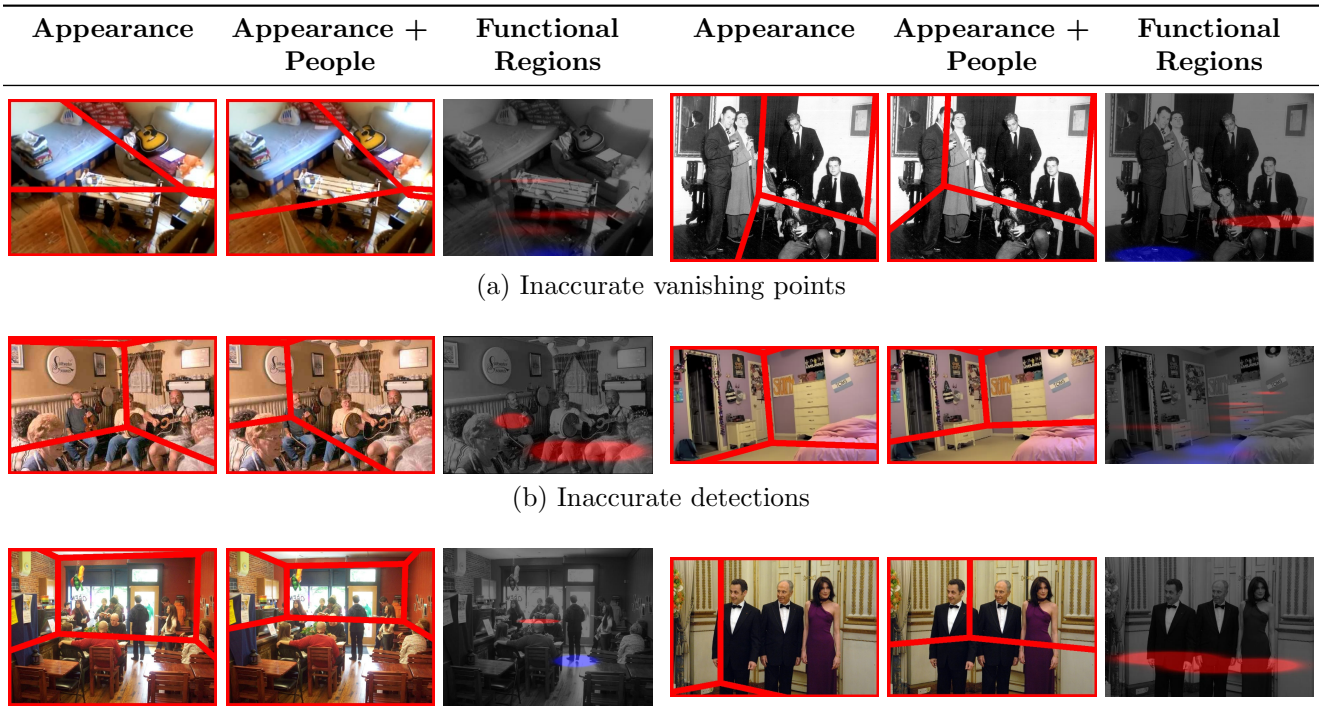


(b) Performance gain v. number of people detected

Fig. 12 Scatterplots of relative performance vs. number of people present and the number of people detected on the extended still image dataset. To better illustrate the density, the x values of points are slightly jittered. Substantial gains ($> 5\%$) are shown in green, approximately equivalent ($\pm 5\%$) in blue, and substantial losses ($< -5\%$) in red (*where substantial losses account for only* 13.7% *of the data*). Our system is able to achieve substantial gains in performance with even a single person acting as a cue.

the number of people detected in Figure 12 (changes in performance with 0 detections are due to our regularization term). Quantitatively, we can analyze the relationship between our relative performance and these two measures of the number of people via Spearman's rank correlation coefficient ρ, which characterizes the degree to which two variables can be related with a monotonic function (i.e., whether increases in the number of people lead to greater gains in performance), with -1 indicating perfect negative rank correlation and 1 perfect rank correlation. Ideally, we would like a weak correlation and a positive one, if any (i.e., ρ near 0 but preferably positive): if there was a strong correlation, then all of our large gains would be the result of the less-likely densely populated scenes. Relative performance is more or less unaffected by the number of people present in the scene ($\rho = 0.0081$, bootstrapped 95% confidence interval -0.0827 to 0.0979). If we consider the number of detections, performance is also more or less unaffected for scenes in which we detect at least one person ($\rho = 0.0564$, bootstrapped 95% confidence interval -0.0416 to 0.1496); there is, of course, a much stronger correlation if we also include scenes with no detections in the analysis since our algorithm requires at least one person to produce large gains.

Accordingly, our system is not limited to scenes in which large numbers of people appear, and crucially can produce substantial gains in sparsely populated scenes which are more likely in consumer photograph collections. To explain this result, we show qualitative examples in Figure 13 of all pairings of small and large performance gains and densely and sparsely filled scenes. When small gains occur, appearance cues give a largely correct interpretation of the scene but cannot disambinguate two interpretations of a small detail (e.g., where the floor ends); in these cases, a person reveals the correct interpretation, producing a better result. Large gains are generally the result of appearance cues being completely incorrect; in these cases, while a large

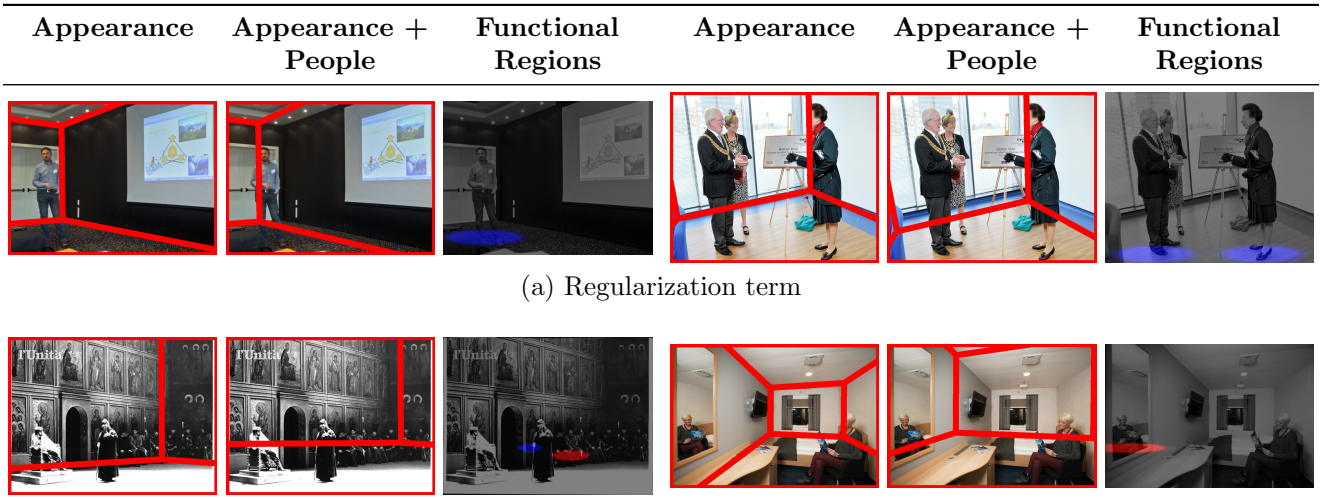


(a) Inaccurate vanishing points

(b) Inaccurate detections

(c) Pose confusion and clipped detections

Fig. 14 Failure modes of the system due to external components. In the single image cases, the variances of the Gaussians for functional surface computation have been increased for clarity and the map has been rendered more strongly.



(a) Regularization term

(b) Limits of model

Fig. 15 Failure modes of the system for both single images and time-lapses intrinsic to our system. In the single image cases, the variances of the Gaussians for functional surface computation have been increased for clarity and the map has been rendered more strongly.

number of people may help provide additional evidence, even a single observation can compensate for the incorrect appearance evidence.

6.3 When does trying to observe people hurt us and why?

Our system is built from noisy detectors, vanishing point estimators, and appearance classifiers and is therefore

by no means perfect. We now present an analysis of our failure modes, which show where future work could perhaps provide improved performance. Here, we focus on modes in which our system produces worse results. In turn, we discuss the dominant causes for our system to have lower performance than the appearance-only system: (a) incorrect vanishing point estimation; (b) inaccurate detections; (c) clipped and occluded detections; (d) regularization (with a caveat); and (e) fundamental limits of our model.

The first three failure modes, illustrated in Figure 14, are due to limitations of components that provide inputs to our system, and can be fixed by improved approaches for their solution or better models learned from better data: our approach is agnostic to the source of vanishing points and, as demonstrated by the use of two detectors, is also agnostic to the source of human poses.

Vanishing point estimation is largely successful on the datasets used, but incorrect estimates that give an incorrect search space can occur, leading to substantial performance losses. For instance, in Figure 14(a) right, a floor has been expanded to agree with functional surfaces; since the hypothesis space is incorrect due to incorrect vanishing points, the floor cannot be expanded correctly and covers the ground-truth wall, making a bad result worse. On the left, the estimation has catastrophically failed, and an already incorrect result is made worse. This is not particular to our approach, but a weakness common to all approaches that search over a single family of vanishing-point aligned cuboidal models.

Incorrect output of the pose estimators can also mislead our system. This can take the straight-forward form of outright false-positives and mistakes in localization that force the room to be expanded incorrectly as in Figure 14(b). This can also happen when the full extent of a person's body is not visible, and the functional contact points are inaccurately predicted, as shown in Figure 14(c): e.g., an occluded standing person predicted as a too-short sitting person or a person with feet below the image predicted with the feet within the image. As future work improves the performance of our subcomponent systems, we anticipate these issues becoming less frequent.

The last failure modes, illustrated in Figure 15 are tied to the proposed model, and cannot be improved by outside algorithmic improvements. A large fraction of our performance losses on a per-image basis can be explained by our regularization term's preference for smaller rooms, illustrated in Figure 15(a); nonetheless, these are generally modest losses, and on average this regularization term improves results and is always made non-zero by parameter learning. Finally, occasionally there are images that violate our model, as shown in Figure 15(b), and thus lead to worse performance. This can happen with scenes with large and populated hallways outside the room, in which case the room is improperly expanded and walls are labeled as floor. Similarly, our system is confused by mirrors or pictures. While edge cases such as mirors and life-sized icons are presently too rare and difficult to handle, the non-cuboidal scene case can be remedied by integrating functional reasoning into other layout estimation models that are not limited to cuboids, such as that of Flint *et al.* [15]: our functional surfaces are not tied to the box-prediction paradigm.

The above discussion describes how the system can have worse performance, but not how it can fail to capitalize on people in the scene. At this point, a large fraction of these cases are explained by the limits of our pose detectors. Another contributing factor is the local nature of our constraints: given an observed contact point between a human and a surface, our system only infers functional properties at that contact point and does not propagate the functional property to similar-appearing regions. In future work, a more complex and learned relationship between human poses and scene properties, such as the one developed in [10], may permit individual joints to provide constraints on room layout that have larger spatial support and higher likelihood of helping correct an inaccurate room estimate.

7 Conclusions

While recognizing actions and estimating poses for a given person is still a very challenging problem, we have shown that noisy pose detections can significantly improve estimates of scene geometry and 3D layout even in a single image. We have shown how even a single person in the right place can greatly improve scene interpretation. In the future, we expect further gains in the accuracy of the proposed method when better pose estimators become available; we further anticipate that our approach to afforance reasoning via functional surfaces can be extended to other models and problems in scene-understanding, for instance non-cuboidal Manhattan-world layout recovery or anomaly detection.

Acknowledgments: This work was supported by NSF Graduate Research and NDSEG Fellowships to DF, and by ONR-MURI N000141010934, NSF IIS-1320083, the MSR-INRIA laboratory, the EIT-ICT labs, Google, ERC Activia, and the Quaero Programme, funded by OSEO.

References

1. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR (2009)
2. Andriluka, M., Roth, S., Schiele, B.: Monocular 3D pose estimation and tracking by detection. In: CVPR (2010)
3. Barinova, O., Lempitsky, V., Tretyak, E., Kohli, P.: Geometric image parsing in man-made environments. In: ECCV (2010)
4. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3D human pose annotations. In: ICCV (2009)
5. Choi, W., Chao, Y.W., Pantofaru, C., Savarese, S.: Understanding indoor scenes using 3D geometric phrases. In: CVPR (2013)
6. Coughlan, J., Yuille, A.: The Manhattan world assumption: Regularities in scene statistics which enable bayesian inference. In: NIPS (2000)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
8. Del Pero, L., Bowdish, J., Fried, D., Kermgard, B., Hartley, E.L., Barnard, K.: Bayesian geometric modeling of indoor scenes. In: CVPR (2012)
9. Del Pero, L., Guan, J., Brau, E., Schlecht, J., Barnard, K.: Sampling bedrooms. In: CVPR (2011)
10. Delaitre, V., Fouhey, D., Laptev, I., Sivic, J., Efros, A., Gupta, A.: Scene semantics from long-term observation of people. In: ECCV (2012)
11. Delaitre, V., Sivic, J., Laptev, I.: Learning person-object interactions for action recognition in still images. In: NIPS (2011)
12. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for static human-object interactions. In: SMiCV, CVPR (2010)
13. Efron, B.: Better bootstrap confidence intervals. Journal of the American Statistical Association **82**(397), 171–185 (1987)
14. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
15. Flint, A., Murray, D., Reid, I.: Manhattan scene understanding using monocular, stereo, and 3D features. In: ICCV (2011)
16. Fouhey, D.F., Delaitre, V., Gupta, A., Efros, A.A., Laptev, I., Sivic, J.: People watching: Human actions as a cue for single-view geometry. In: ECCV (2012)
17. Fouhey, D.F., Gupta, A., Hebert, M.: Data-driven 3D primitives for single image understanding. In: ICCV (2013)
18. Gall, J., Fossati, A., van Gool, L.: Functional categorization of objects using real-time markerless motion capture. In: CVPR (2011)
19. Gibson, J.: The ecological approach to visual perception. Boston: Houghton Mifflin (1979)
20. Grabner, H., Gall, J., van Gool, L.: What makes a chair a chair? In: CVPR (2011)
21. Guan, L., Franco, J.S., Pollefeys, M.: 3D occlusion inference from silhouette cues. In: CVPR (2007)
22. Gupta, A., Chen, T., Chen, F., Kimber, D., Davis, L.: Context and observation driven latent variable model for human pose estimation. In: CVPR (2008)
23. Gupta, A., Davis, L.S.: Objects in action: An approach for combining action understanding and object perception. In: CVPR (2007)
24. Gupta, A., Efros, A., Hebert, M.: Blocks world revisited: Image understanding using qualitative geometry and mechanics. In: ECCV (2010)
25. Gupta, A., Satkin, S., Efros, A., Hebert, M.: From 3D scene geometry to human workspace. In: CVPR (2011)
26. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, second edn. Cambridge University Press, ISBN: 0521540518 (2004)
27. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: ICCV (2009)
28. Hedau, V., Hoiem, D., Forsyth, D.: Thinking inside the box: Using appearance models and context based on room geometry. In: ECCV (2010)
29. Hoiem, D., Efros, A., Hebert, M.: Geometric context from a single image. In: ICCV (2005)
30. Hoiem, D., Efros, A., Hebert, M.: Putting objects in perspective. IJCV (2008)
31. Jiang, Y., Saxena, A.: Hallucinated humans as the hidden context for labeling 3D scenes. In: CVPR (2013)
32. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: CVPR (2011)
33. Kanade, T.: Recovery of the three-dimensional shape of an object from a single view. Artificial Intelligence **17**(1), 409–460 (1981)
34. Karsch, K., Liu, C., Kang, S.B.: Depth extraction from video using non-parametric sampling. In: ECCV (2012)
35. Kjellstrom, H., Romero, J., Martinez, D., Kragic, D.: Simultaneous visual recognition of manipulation actions and manipulated objects. In: ECCV (2008)
36. Krahnstoever, N., Mendonca, P.R.S.: Bayesian autocalibration for surveillance. In: CVPR (2005)
37. Lee, D., Gupta, A., Hebert, M., Kanade, T.: Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In: NIPS (2010)
38. Lee, D., Hebert, M., Kanade, T.: Geometric reasoning for single image structure recovery. In: ICCV (2009)
39. Park, D., Ramanan, D.: N-best maximal decoders for part models. In: ICCV (2011)
40. Payet, N., Todorovic, S.: Scene shape from texture of objects. In: CVPR (2011)
41. Prest, A., Schmid, C., Ferrari, V.: Weakly supervised learning of interactions between humans and objects. PAMI (2011)
42. Ramakrishna, V., Kanade, T., Sheikh, Y.: Tracking human pose by tracking symmetric parts. In: CVPR (2013)
43. Rother, C.: A new approach to vanishing point detection in architectural environments. IVC **20** (2002)
44. Rother, D., Patwardhan, K., Sapiro, G.: What can casual walkers tell us about the 3D scene. In: CVPR (2007)
45. Saxena, A., Sun, M., Ng, A.Y.: Make3D: Learning 3D scene structure from a single still image. TPAMI (2008)
46. Schodl, A., Essa, I.: Depth layers from occlusions. In: CVPR (2001)
47. Schwing, A.G., Fidler, S., Pollefeys, M., Urtasun, R.: Box in the box: Joint 3D layout and object reasoning from single images. In: ICCV (2013)
48. Schwing, A.G., Urtasun, R.: Efficient Exact Inference for 3D Indoor Scene Understanding. In: ECCV (2012)
49. Taylor, C.J.: Reconstruction of articulated objects from point correspondences in a single image. In: CVPR (2000)
50. Turek, M., Hoogs, A., Collins, R.: Unsupervised learning of functional categories in video scenes. In: ECCV (2010)
51. Wang, H., Gould, S., Koller, D.: Discriminative learning with latent variables for cluttered indoor scene understanding,. In: ECCV (2010)

52. Xiao, J., Russell, B., Torralba, A.: Localizing 3D cuboids in single-view images. In: NIPS (2012)
53. Yang, Y., Ramanan, D.: Articulated pose estimation using flexible mixtures of parts. In: CVPR (2011)
54. Yao, B., Khosla, A., Fei-Fei, L.: Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses. In: Proc. ICML (2011)
55. Yu, S.X., Zhang, H., Malik, J.: Inferring spatial layout from a single image via depth-ordered grouping. In: 6th Workshop on Perceptual Organization in Computer Vision (2008)