

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

▶ **To cite this version:**

David F. Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A. Efros, Ivan Laptev, et al.. People Watching: Human Actions as a Cue for Single View Geometry. ECCV'12 - 12th European Conference on Computer Vision, Oct 2012, Florence, Italy. pp.732-735, 10.1007/978-3-642-33715-4_53 . hal-01060874

HAL Id: hal-01060874
https://inria.hal.science/hal-01060874

Submitted on 4 Sep 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

People Watching: Human Actions as a Cue for Single View Geometry

David F. Fouhey[1], Vincent Delaitre[2],
Abhinav Gupta[1], Alexei A. Efros[1,2], Ivan Laptev[2], and Josef Sivic[2]

[1]Carnegie Mellon University [2]INRIA/École Normale Supérieure, Paris
http://graphics.cs.cmu.edu/projects/peopleWatching/

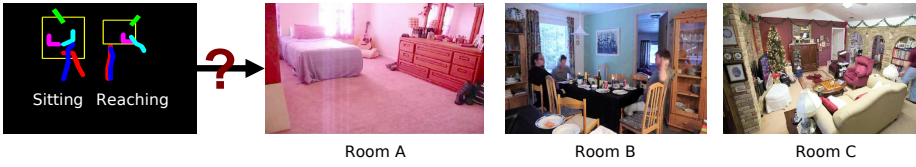


Fig. 1. What can human actions tell us about the 3D structure of a scene? Quite a lot, actually. Consider the people depicted on the left. They were detected in a time-lapse sequence in one of rooms A, B, or C. Which room did they come from? See the text for the answer.

Abstract. We present an approach which exploits the coupling between human actions and scene geometry. We investigate the use of 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 about the scene. These constraints are then used to improve state-of-the-art single-view 3D scene understanding approaches. The proposed method is validated on a collection of monocular time-lapse sequences collected from YouTube and a dataset of still images of indoor scenes. We demonstrate that observing people performing different actions can significantly improve estimates of 3D scene geometry.

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

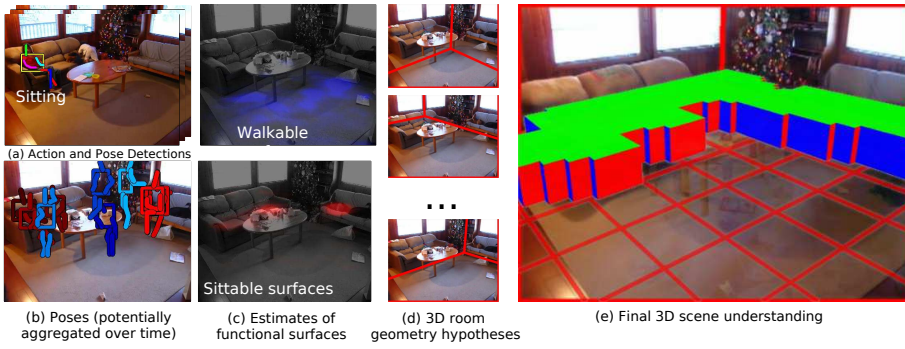


Fig. 2. Overview of the proposed approach. We propose the use of both appearance and human action cues for estimating single-view geometry. Given an input image or set of input images taken by a fixed camera (e.g., a time-lapse), we estimate human poses in each image (a), yielding a set of human-scene interactions (b), which we aggregate over time (for time-lapses). We use these to to infer functional surfaces (c) in the scene: sittable (red), walkable (blue). We simultaneously generate multiple 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. **See all results on the project website.**

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 represent human poses following J.J. Gibson's notion of *affordances* [1] – 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. 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 [2–4], and it is possible to use anthropometric constraints to lift the poses into 3D [5], the accuracy of these methods is still too low to be used reliably.

As a result, in this paper we take a soft, hybrid approach. We first employ the single-view indoor reconstruction method of Hedau *et al.* [6] 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.

1.1 Background

Our goal is to understand images in terms of 3D geometry and space. Traditional approaches in computer vision have focused on using correspondences and multiple view geometry for 3D reconstruction. While these methods have been successful, they are not applicable when only a single view of the scene is available. Since humans can infer scene structure from a single image, single-view reconstruction is a critical step towards vision systems with more 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 tasks such as object detection [7, 8].

In recent years, there have been significant advances in using statistical approaches for single-view 3D image understanding [6, 9–18]. To make progress on the extremely difficult and severely underconstrained problem of estimating scene geometry from a single image, these approaches impose domain specific constraints, mainly regarding the human-made nature of the scenes. However, although they assume a human-centric scene structure, each of these approaches treats humans as clutter rather than as a cue. 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 [19–21]. 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 [20, 22, 23], action recognition [19, 21, 24, 25], and pose estimation [26, 27].

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 [28, 29]. For instance, Gupta *et al.* [29] argued that functional questions such as "Where can I sit?" are more important than categorizing objects based on name, and used estimated 3D geometry in images to infer Gibsonian affordances [1], or "opportunities for interaction" with the environment. Our work focuses on the inverse of the problem addressed in [28, 29]: 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 [2–4, 30, 31], and design a method to improve single-view indoor geometry estimation. Even though the building blocks of this work, human pose estimation [3] and 3D image understanding [6, 10], 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 silhouttes [32] in multi-view setups. In single-view scenarios, the focus has been on coarse constraints from person tracks [33–35], whereas we focus on fine-grained physical and functional constraints using human actions and poses.

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 approaches exist specifically for estimating indoor scene geometry [6, 10, 8]. 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 scene layout estimation [6, 10]. Specifically, we build upon the work of Hedau *et al.* [6]: 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 (Manhattan world [36]) and therefore estimating a room involves fitting a parametric 3D box.

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. This type of data is ideal for our experiments since it has a high diversity of person-scene interactions. It also enables us to test our method on realistic data with non-staged activities in a variety of natural environments.

The 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 [6] to predict the global 3D geometry of the room by selecting the best hypothesis from a set of possible hypotheses

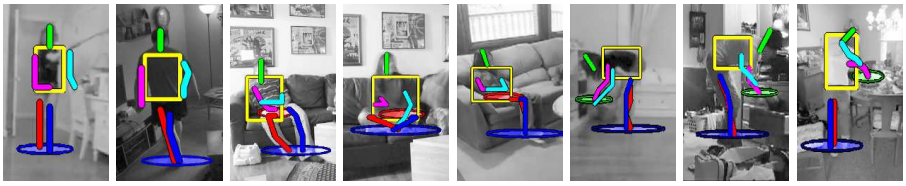


Fig. 3. Example action detection and pose estimation results. The predicted surface contact points are shown by ellipses: blue (walkable), red (sittable), green (reachable). Shown actions are: standing (1-2), sitting (3-5), and reaching (6-8).

(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 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.

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., beds are sittable and reachable). To increase robustness, we place a Gaussian at the contact points of each detection and weight the contribution of the pose by the classifier 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. This yields probability maps h for the different types of functional image regions, as illustrated in Figures 2c and 4c,d.

Our approach is agnostic to the particular method of pose detection; in this work, we use two complementary approaches. We build primarily on the articulated pose model of Yang and Ramanan [3]. Here, we employ the model for detecting human action by training a separate model for each of the three actions. Additionally, we use the model of Felzenszwalb *et al.* [31] for sitting and standing: the low variance of the relevant joints of these actions (e.g., feet for standing) enable us to accurately approximate poses by simply transferring a fixed pose. Since the sitting detector may also respond to standing people, we discriminate between different actions by jointly calibrating the detectors of each model with respect to each other by fitting a multinomial logistic regression



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

Fig. 4. 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.

model. Action classification is performed in a non-maxima suppression manner: if bounding boxes of several detections overlap, then the detection with the highest calibrated response is kept. The articulated pose estimator and deformable parts model are calibrated separately, and produce independent estimates of functional regions. Examples of detected actions together with estimated body pose configurations and predicted contact points are shown in Figure 3.

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 [29], 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, 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. 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.* [6] which first estimates three orthogonal vanishing points and then samples multiple room layout hypotheses that are aligned with the estimated three scene directions. The best hypothesis is selected using a learned scoring function based on global appearance cues, such as detected straight lines and classifier outputs for different surfaces (walls, floor, ceiling). 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. 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 [6], and 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), our goal is to find the best room layout hypothesis y^*. We use the following scoring function to evaluate the coherence of image features and human poses with the hypothesized room layout y:

$$f(x, H, y) = \psi(x, y) + \phi(H, y) + \rho(y), \tag{1}$$

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.

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 [6]. The second term enforces the containment constraints and expands as

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

where $\zeta(h)$ is the mapping of support 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 both the human body and the objects it is interacting with should lie inside the room. We approximate $\zeta(h)$ by using the feet locations of detected actors, which produces accurate results for our action vocabulary. Finally, the term $\rho(y) = -c \cdot \max(0, (A - M)/M)$ imposes a penalty for excessive floor area A, measured with respect to the minimum floor area M out of the top three hypotheses; in our experiments, $c = 1/8$. We need this regularization term since $\phi(H, y)$ can only expand the room to satisfy the containment constraint.

4.2 Estimating Free Space in the Scene

Once we have estimated room layouts we now estimate the voxels occupied by objects. However, this is a difficult and ill-posed problem. Hedau *et al.* [6] 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.* [10] and Gupta *et al.* [29] 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.* [6], and then incorporate constraints from different human poses to further refine this occupied voxel map. Specifically, we backproject each functional image region h at its 3D height[1], yielding a horizontal slice inside the voxel map. 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. The final score for occupancy of a particular voxel is a linear sum of these votes, weighed by the confidence of human pose detections; as the result is probabilistic, to produce a binary interpretation, we must threshold the results.

5 Experiments

We validate the proposed approach on a collection of consumer time-lapse videos of indoor scenes and a collection of indoor still images. Both datasets were collected from the Internet and depict challenging scenes capturing one or more people engaged in everyday activities interacting with the scene. The code and data for the experiments is available on the project webpage.

Baselines. For both time-lapses and single images, we compare our estimates of room geometry to a number of approaches. Our primary baseline is the appearance-only system of Hedau *et al.* [6]. To provide context, we also include another baseline, in which we impose the box model on the approach of Lee *et al.* [37]. Finally, to show that all methods are operating better than chance, we use location only to predict the pixel labels: after resizing all scenes to common dimensions, we use the majority label in the training images for each pixel.

We use the standard metric of per-pixel accuracy. We compare the estimated layout with a manual labeling of room geometry; note that since the camera is fixed in time-lapses, only a single annotation is needed.

Implementation details. We train detectors on example images using the Yang and Ramanan model for all three actions [3] and the Felzenszwalb *et al.*

[1] Because our classes are fine-grained, we can use human dimensions for the heights of the surfaces: for reaching, it is waist height (3ft), and sitting, knee-height (1ft).

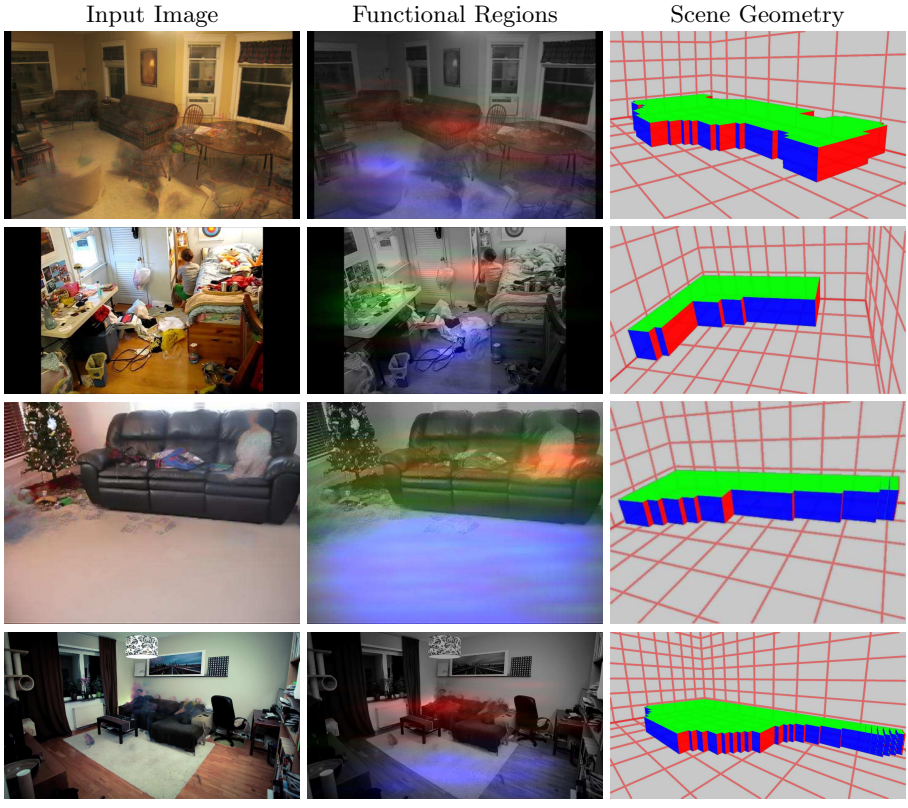
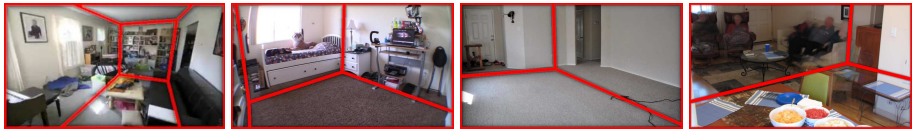
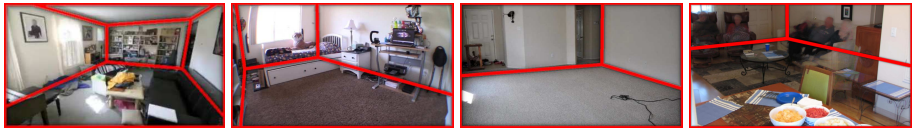


Fig. 5. Example time-lapse sequence results: given an input image, we use functional regions (walkable: blue; sittable: red; reachable: green) to constrain the room layout; having selected a layout, we can also infer a more fine-grained geometry of the room via functional reasoning. **See all results on the project website.**



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



(b) Appearances + People (Our approach).

Fig. 6. Timelapse experiment: A comparison of (a) appearance only baseline [6] with (b) our improved room layout estimates. In many cases, the baseline system selects small rooms due to high clutter. On the right, even though the room is not precisely a cuboid, our approach is able to produce a significantly better interpretation of the scene.

Table 1. Time-lapse experiment: Average pixel accuracy for geometry estimation on time-lapse sequences. Our method achieves significant gains; further, using humans alone produces competitive performance.

	Location	**Appearance Only**		**People Only**	**Appearance + People**
		Lee *et al.*	Hedau *et al.*		
Overall	64.1%	70.4 %	74.9%	70.8%	**82.5%**

model for sitting and standing. For the standing action, we use a subset of 196 images from [3] 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 [38], indoor scenes from [6], 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 and remove false-positive detections on the background. We also use geometric filtering similar to [7] to remove detections that significantly violate the assumption of single ground plane.

5.1 Experiment 1: Understanding time-lapse sequences

This dataset contains 40 videos of indoor time-lapse sequences totaling more than 140,000 frames. The videos were collected from YouTube by using keywords such as "time-lapse," "living room," "dinner," "party," or "cleaning." We treat each sequence as a collection of still images of a particular scene. Most of the frames contain one or more people interacting with the scene or each other. Examples include: people sitting on beds or sofas; people walking and people reaching into drawers and on tables. On average each video has around 3500 frames and 1200, 1300 and 400 detections of standing, sitting and reaching action, respectively.

Figure 5 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 first case, the far away couches and their corresponding sittable surfaces are hard to recognize, even for human observers. Because our approach observes human actors walking and sitting in those areas, it can easily infer the sittable surface. Similarly, 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.* [6]. Figure 6 shows some examples of the relative performance; comparisons for all time-lapse sequences may be found on the project page. Quantitatively, as shown in Table 1, our method is able to consistently improve on the baseline, averaging a gain of 7.6% (bootstrapped 95% confidence

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
Overall	66.4%	71.3%	77.0%	79.6%	**80.8%**

interval: 4.5% to 11.3%). Further, our performance is worse than the baseline in only 7.5% of cases. To demonstrate the value of cues from people, we show results only using human action cues to select room hypotheses; specifically, we use only our human action compatibility and room size terms, ϕ and ρ, to rank the hypotheses. Even with only people as cues, our system performs only 4.1% worse on average than Hedau *et al.* and equivalently to Lee *et al.*

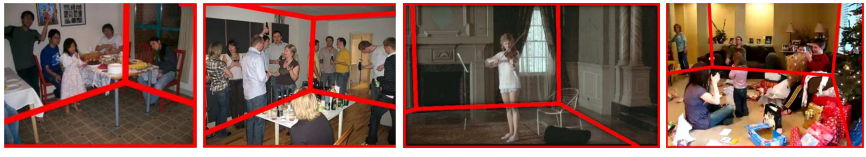
Following [39], we also quantitatively evaluate estimated free space. We obtain ground-truth by manually labeling the floor occupancy map in the estimated room. Compared to the appearance-only backprojected clutter labeling, our approach achieves a 15.1% average precision gain in estimating floor free space.

5.2 Experiment 2: Understanding single images

In the second experiment, we consider a dataset of 100 still images of indoor scenes. As existing work treats humans as clutter, previous data sets have deliberately excluded humans from their scenes; we therefore must gather a new dataset. The images were collected with Internet image search engines using keywords such as "living room," "eating," or "sitting," and in collections of pictures of celebrities and political figures. We emphasize that since our approach is general, it can be applied to the wealth of still images available on the Internet.



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



(b) Appearances + People (Our approach)

Fig. 7. Still Image Experiments: The correct person in the correct place can very easily disambiguate complex scene interpretation problems. In the last example, although the vanishing points are inaccurate, we produce a more accurate intepretation.

Our results show that functional constraints from human actions provide strong evidence of 3D geometry even in a single image. Figure 7 shows few examples of our estimated room geometry as compared to Hedau *et al.* [6]. Comparisons for all images may be found on the project website. Figure 8 shows examples of estimated 3D room geometry and the 3D occupied voxels. Quantitatively, we demonstrate 2.6% improvement (bootstrapped 95% confidence interval: $0.9\% - 4.8\%$) over all still images, as seen in Table 2, and our performance is as good or better than [6] in 88% of cases. The gain over the baseline is lower in the still image case than in the time-lapse case; this is largely due to rooms in which functional reasoning does not significantly adjust the interpretation, leading to equivalent accuracy with and without people: in many cases, human actions cannot be exploited, even if the person detections are perfect (Table 2), e.g., if the room selected with appearance alone is correct or if all actors are contained within an inaccurate room.

6 Discussion

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. This suggest other ways of using statistically aggregated noisy pose estimates, for example, learning relations between human actions and semantic objects (beds, chairs, tables) in the scene [40]. We expect further gains in accuracy of the proposed method when better pose estimators become available in future.

Acknowledgments: This work was supported by a NSF Graduate Research Fellowship to DF, and by ONR-MURI N000141010934, Quaero, OSEO, MSR-INRIA, EIT-ICT, and ERC grant Videoworld.

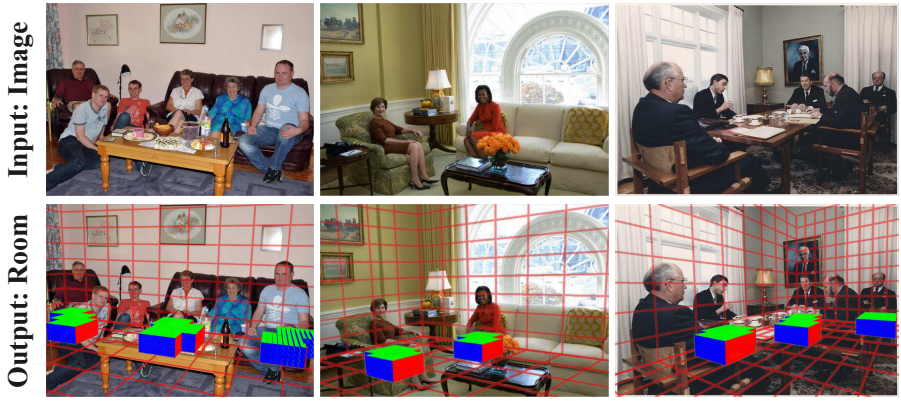


Fig. 8. Still Image Experiments: Functional reasoning to detect sittable surfaces in still images.

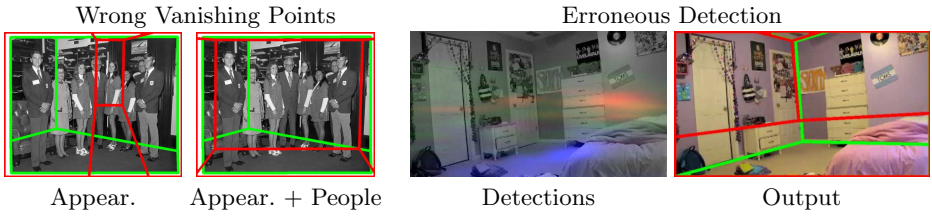


Fig. 9. Failure cases (ground-truth room in green): in some cases, vanishing point extraction fails due to clutter. In other cases, there are plausible but inaccurate detections.

References

1. Gibson, J.: The ecological approach to visual perception. Boston: Houghton Mifflin (1979)
2. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR. (2009)
3. Yang, Y., Ramanan, D.: Articulated pose estimation using flexible mixtures of parts. In: CVPR. (2011)
4. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: CVPR. (2011)
5. Taylor, C.J.: Reconstruction of articulated objects from point correspondences in a single image. In: CVPR. (2000)
6. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: ICCV. (2009)
7. Hoiem, D., Efros, A., Hebert, M.: Putting objects in perspective. IJCV (2008)
8. Hedau, V., Hoiem, D., Forsyth, D.: Thinking inside the box: Using appearance models and context based on room geometry. In: ECCV. (2010)
9. Yu, S.X., Zhang, H., Malik, J.: Inferring spatial layout from a single image via depth-ordered grouping. In: The 6th IEEE Computer Society Workshop on Perceptual Organization in Computer Vision. (2008)
10. Lee, D., Gupta, A., Hebert, M., Kanade, T.: Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In: NIPS. (2010)
11. Hoiem, D., Efros, A., Hebert, M.: Geometric context from a single image. In: ICCV. (2005)
12. Wang, H., Gould, S., Koller, D.: Discriminative learning with latent variables for cluttered indoor scene understanding,. In: ECCV. (2010)
13. Gupta, A., Efros, A., Hebert, M.: Blocks world revisited: Image understanding using qualitative geometry and mechanics. In: ECCV. (2010)
14. Barinova, O., Lempitsky, V., Tretyak, E., Kohli, P.: Geometric image parsing in man-made environments. In: ECCV. (2010)
15. Del Pero, L., Guan, J., Brau, E., Schlecht, J., Barnard, K.: Sampling bedrooms. In: CVPR. (2011)
16. Payet, N., Todorovic, S.: Scene shape from texture of objects. In: CVPR. (2011)
17. Schwing, A., Hazan, T., Pollefeys, M., Urtasun, R.: Efficient structured prediction for 3D indoor scene understanding. In: CVPR. (2012)

18. Del Pero, L., Bowdish, J., Fried, D., Kermgard, B., Hartley, E.L., Barnard, K.: Bayesian geometric modeling of indoor scenes. In: CVPR. (2012)
19. Gupta, A., Davis, L.S.: Objects in action: An approach for combining action understanding and object perception. In: CVPR. (2007)
20. Turek, M., Hoogs, A., Collins, R.: Unsupervised learning of functional categories in video scenes. In: ECCV. (2010)
21. Delaitre, V., Sivic, J., Laptev, I.: Learning person-object interactions for action recognition in still images. In: NIPS. (2011)
22. Prest, A., Schmid, C., Ferrari, V.: Weakly supervised learning of interactions between humans and objects. PAMI (2011)
23. Gall, J., Fossati, A., van Gool, L.: Functional categorization of objects using real-time markerless motion capture. In: CVPR. (2011)
24. Kjellstrom, H., Romero, J., Martinez, D., Kragic, D.: Simultaneous visual recognition of manipulation actions and manipulated objects. In: ECCV. (2008)
25. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for static human-object interactions. In: SMiCV, CVPR. (2010)
26. 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)
27. Gupta, A., Chen, T., Chen, F., Kimber, D., Davis, L.: Context and observation driven latent variable model for human pose estimation. In: CVPR. (2008)
28. Grabner, H., Gall, J., van Gool, L.: What makes a chair a chair? In: CVPR. (2011)
29. Gupta, A., Satkin, S., Efros, A., Hebert, M.: From 3d scene geometry to human workspace. In: CVPR. (2011)
30. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3D human pose annotations. In: ICCV. (2009)
31. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR. (2008)
32. Guan, L., Franco, J.S., Pollefeys, M.: 3d occlusion inference from silhouette cues. In: CVPR. (2007)
33. Krahnstoever, N., Mendonca, P.R.S.: Bayesian autocalibration for surveillance. In: CVPR. (2005)
34. Rother, D., Patwardhan, K., Sapiro, G.: What can casual walkers tell us about the 3D scene. In: CVPR. (2007)
35. Schodl, A., Essa, I.: Depth layers from occlusions. In: CVPR. (2001)
36. Coughlan, J., Yuille, A.: The Manhattan world assumption: Regularities in scene statistics which enable bayesian inference. In: NIPS. (2000)
37. Lee, D., Hebert, M., Kanade, T.: Geometric reasoning for single image structure recovery. In: ICCV. (2009)
38. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
39. Hedau, V., Hoiem, D., Forsyth, D.: Recovering free space of indoor scenes from a single image. In: CVPR. (2012)
40. Delaitre, V., Fouhey, D., Laptev, I., Sivic, J., Efros, A., Gupta, A.: Scene semantics from long-term observation of people. In: ECCV. (2012)