Towards Complete Scene Reconstruction from Single-View Depth and Human Motion

Sam Fowler
sam.fowler@surrey.ac.uk
Hansung Kim
h.kim@surrey.ac.uk
Adrian Hilton
a.hilton@surrey.ac.uk

Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, UK

cvssp.org/projects/s3a/CompRecon

Abstract

Complete scene reconstruction from single view RGBD is a challenging task, requiring estimation of scene regions occluded from the captured depth surface. We propose that scene-centric analysis of human motion within an indoor scene can reveal fully occluded objects and provide functional cues to enhance scene understanding tasks. Captured skeletal joint positions of humans, utilised as naturally exploring active sensors, are projected into a human-scene motion representation. Inherent body occupancy is leveraged to carve a volumetric scene occupancy map initialised from captured depth, revealing a more complete voxel representation of the scene. To obtain a structured box model representation of the scene, we introduce unique terms to an object detection optimisation that overcome depth occlusions whilst deriving from the same depth data. The method is evaluated on challenging indoor scenes with multiple occluding objects such as tables and chairs. Evaluation shows that human-centric scene analysis can be applied to effectively enhance state-of-the-art scene understanding approaches, resulting in a more complete representation than single view depth alone.

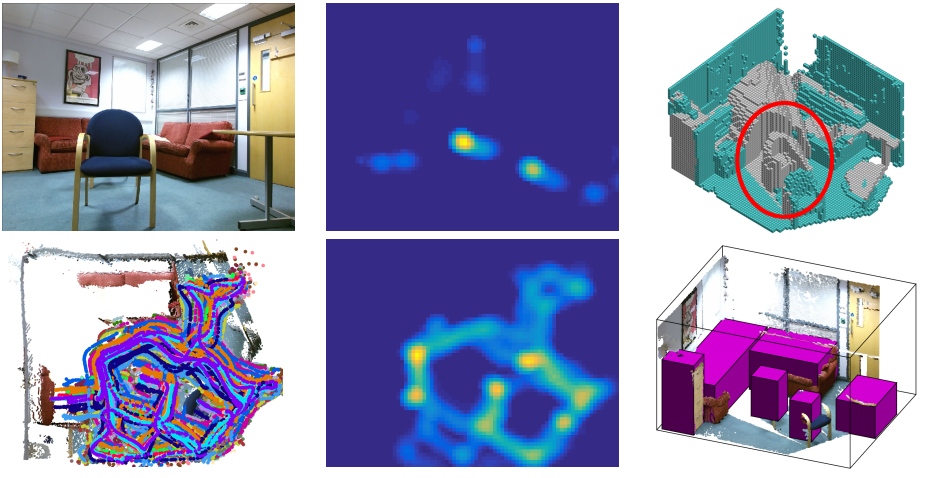
1 Introduction

The introduction of commodity depth sensors to the market has had a profound impact on computer vision. Indoor scene understanding has benefited greatly from the high precision, low range sensors that have proven to improve upon monocular benchmarks across detection and reconstruction tasks. Depth sensing has also advanced the accuracy of human skeletal tracking, providing precise localisation of 3D joint positions within a captured scene.

Famously, KinectFusion [18] introduced real-time surface mapping by tracking a single mobile sensor. The moving sensor system allows large regions of indoor scenes to be scanned, requiring users to move the sensor to produce complete scene reconstructions. Fixed sensor systems, whilst requiring no user interaction, suffer the common issues of limited field of view and both self- and scene-occlusion, limiting their capability to reconstruct the complete scene.

In this paper, we investigate how analysis of human motion within a scene can contribute to scene understanding tasks, with particular interest in occluded object localisation.

© 2017. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.



(a) (b) (c)

Figure 1: **Proposed system:** (a) Long-term capture of the scene, providing RGBD frames of the scene whilst static (top) and 3D skeleton joint positions within the scene when dynamic (bot.); (b) Skeleton joint data is projected into an overhead representation of the scene. Human poses for sitting and standing are recognised to create motion maps for functionably afforded sitting (top) and standing (bot.) scene regions; (c) Motion maps are applied to carve a volumetric occupancy map which was initialised on captured depth data (top; blue voxels represent captured depth, grey voxels represent occluded scene regions). A linear optimisation utilises the 3D occupancy map and 2D motion maps to detect visible and occluded objects (bot.). Notice how an occluded chair is detected behind the visible chair.

Overcoming the process of scanning a scene for a complete reconstruction, we as humans naturally reveal the extent of an indoor environment when we utilise it in our day-to-day lives. Tracking the movement of a body behind an object can reveal hidden free space and recognition of humans actions such as sitting locations can aid scene understanding and object detection; we explore how these can be utilised in this work.

The proposed system produces a probability map of human motion within the scene, from an overhead perspective after aligning the scene to a Manhattan-world coordinate system. After recognising human poses to separate the motion map to represent likely scene regions for sitting and standing, a scene voxel occupancy map initialised with free and occluded space from depth data is carved to give a more complete representation of the scene. This representation allows hidden regions to be realised through the human motion analysis. A linear integer programming system is applied to detect aligned boxes for visible and occluded objects, utilising the occupancy and motion maps to eliminate the occlusion issues characteristic to depth data. Figure 1 depicts an overview of the approach.

The key contributions of this work are: (1) Manifesting human motion data into a scene-centric representation to expose novel scene cues; (2) Coupling depth and human motion data into a volumetric complete scene occupancy map to reveal occluded scene regions; (3) Detecting fully occluded objects by applying human motion to improve object localisation.

2 Related Work

We study work related to indoor scene reconstruction, considering 3D reconstruction and scene completion approaches, and examining the applications of human-scene interaction.

3D Scene Reconstruction. Many works have studied geometric [1, 3] and structured 3D reconstruction [11, 16, 19, 24]. Geometric approaches such as KinectFusion [18] must scan the scene to produce a complete reconstruction, requiring user interaction to manoeuvre the sensor. Whilst recovering the geometric shape of the scene, objects are not localised and no semantic information is produced. Structured systems utilise geometric primitives such as planes and cuboids, and recently object mesh models [6], to recover semantically meaningful 3D reconstructions. However, these approaches perform on static sensor input, providing only the visible surface to match scene geometry to; in cluttered scenes, occlusions are common and can mask the full extent of visible scene objects and completely obscure areas of scene geometry.

Scene Completion. Interest in reconstruction of unobserved geometry has induced work investigating scene extent completion through planar surface completion [8, 23] and supervised volumetric estimation models [9, 25]. The former works utilise spatial and 3D structure priors to infer the extent that planes are extended beyond observed data. These approaches suffer when planar surfaces are hard to detect due to heavy occlusion, small sensor to plane angle or being above camera height. The voxel completion approaches leverage trained models to map observed data to unobserved scene regions. Song *et al*. [25] couple the often separated tasks of scene completion and semantic labelling to employ the object extent information implicit in object recognition. This supplies precise voxel predictions around visible objects, but can suffer from confusion between similar objects and does not consider regions of complete occlusion.

Human-Scene Interaction. Analysing human movement within a scene can be applied to both recognising scene object affordance [7, 12, 13, 15, 26] and to improve semantic understanding tasks [2, 5]. For the former, Gupta *et al*. [10] hallucinate human poses to localise human action areas, introducing the constraints that the 3D space occupied by a pose cannot intersect scene structure, and the surfaces required to support a pose must be available. Few works utilise human-scene interactions toward semantic scene understanding. With interest in 3D reconstruction, Fouhey *et al*. [5] monitor human motion in time-lapse video to improve room layout estimation, reconstructing a coarse and incomplete estimation of 3D scene geometry.

Overview. Scene reconstruction from single-view depth suffers from complete-, partial- and self-occlusion constraints inherent to the depth surface. Current 3D reconstruction approaches naturally estimate self-occlusions through model fitting, and unobserved region completion frameworks are pursuing occlusion estimation using trained models. This work investigates how cues from human motion, captured in single-view depth sequences, can transform depth into a representation that overcomes these occlusion constraints, allowing reconstruction of fully unobserved scene regions to enhance scene understanding tasks.

3 Detecting Unobserved Scene Regions

This section describes our approach to fitting aligned 3D object bounding boxes to the volumetric occupancy map produced from depth capture and analysis of human motion. Utilising human motion supports the use of the complete scene occupancy map, overcoming the oc-

clusion faced when using depth from static sensor captures. Section 3.1 details how human motion is combined with static depth data to refine a volumetric occupancy map, describing some of the unique data this representation provides. A linear object detection system designed to utilise the occupancy map to overcome scene occlusions is then described in Section 3.2.

3.1 Scene Occupancy map

The occupancy map is formulated as a 3D matrix, size $X \times Y \times Z$, of voxels $V_{i,j,k}$, $i = 1,...,X$, $j = 1,...,Y$, $k = 1,...,Z$ size ${S_v}^3$, where $S_v = 0.05m$ in this paper. This matrix represents the complete volume of the captured room; each voxel $V_{i,j,k}$ is binary, describing the unoccupied and occupied regions of the scene. RGBD capture of an indoor scene is aligned with a Manhattan axis [22] and a room layout estimation technique [9] is applied to predict the outer bounds of the indoor scene. Rounded to S_v, the room bounds determine the number of voxels in each of the scene directions X, Y, Z.

The occupancy map is initialised with the captured depth when the scene is static: voxels between the sensor and the depth surface are known to be unoccupied and so for these, $V = 0$; voxels containing captured depth points are known to be occupied and so $V = 1$. It is unknown whether the remaining voxels, those occluded by the depth surface, are occupied or not: we initialise these to occupied, $V = 1$, so that free space can be revealed by human motion.

3.1.1 Motion Maps

Humans active within the scene are tracked with [21] to provide 3D joint positions over long-term observations of their movement. The use of a machine learning based full-body pose estimation from observed depth allows joint positions to be estimated for occluded body parts. However, this estimation is imprecise, producing inaccurate occluded joint localisations that jitter between frames.

In cluttered indoor scenes, body extremities are often occluded by scene objects and the camera field of view. We hypothesise that the reliably localised larger upper body parts can be applied with body shape constraints to reconstruct these occluded regions. This approach to revealing hidden scene structure allows refinement of the complete scene reconstruction using partially visible body parts.

The considered joints - the head, neck, shoulders, back and hips - are denoted J. Throughout capture, joint positions P_J are projected onto a 2D probabilistic ground-plane projection of the scene, motion map M_{motion}. The average position P_{xyz} of P_J is taken in each frame and a 2D Gaussian cumulatively applied at P_{xz} on the $X \times Z$ sized M_{motion}.

To determine body supporting surfaces in the complete scene, semantics of the human are introduced to the motion map representation. Key body poses are recognised by empirically-trained models of characteristic skeletal joint positions, and pose co-occurrence is recorded over time. In the current implementation, these poses correspond to sitting and standing/walking to provide semantic information on functional support surfaces in the corresponding scene regions. The $X \times Z$ sized binary masks B_{sit} and B_{stand} record pose co-occurrence, resolving the support type of each region. By masking M_{motion}, two motion maps M_{sit} and M_{stand} are created; each pixel (i,k) of M_{sit} and M_{stand} then represents the likelihood that a person has sat or stood within the corresponding vertical column of voxels $V_{i,k}$, and describes the required functional support.

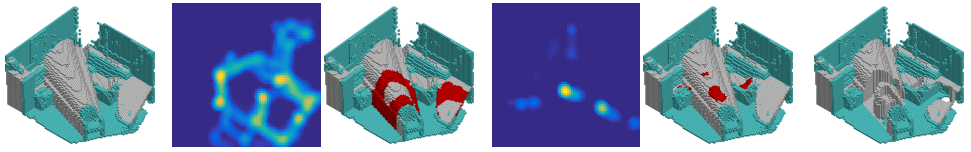


Figure 2: **Volumetric occupancy carving.** The volumetric occupancy map is initialised on captured depth before human motion maps M_{stand} and M_{sit} are applied with Eq. 1 to carve the complete scene. Blue voxels represent captured depth, grey voxels represent occluded regions and red voxels represent voxels removed by carving.

3.1.2 Scene Carving

As visualised in Figure 2, recognised human motion is applied to modify the 3D occupancy map $V_{i,j,k}$ to resolve the occupancy of occluded scene regions. Motion maps M_{sit} and M_{stand} are thresholded such that likelihoods greater than T represent scene regions that have been occupied by sitting or standing poses. Body occupancy constraints are defined as: (i) for regions where standing poses are detected, the space between the standing surface height H_{floor} and the human height $H_{standing}$ is assumed to be free, and (ii) for regions where sitting poses are detected, the space between the sitting surface height H_{seat} and the corresponding human height $H_{sitting}$ is assumed to be free. All heights are derived from the recognised poses, and in this paper $T = 0.1$. This definition allows non-occluded body parts to determine occluded scene regions, and is applied as follows to carve the scene:

$$\begin{aligned} V_{i,j,k} &= 0, \forall\{i,j,k\} \text{ that } M_{stand}(i,k) > T \text{ and } H_{floor} \leq j \leq H_{standing}, \\ V_{i,j,k} &= 0, \forall\{i,j,k\} \text{ that } M_{sit}(i,k) > T \text{ and } H_{seat} \leq j \leq H_{sitting}. \end{aligned} \tag{1}$$

3.2 Scene Representation by Object Bounding Box Detection

Object detection is performed with an integer programming optimisation inspired by [11]; a new optimisation function is introduced to apply the carved volumetric occupancy map to complete scene reconstruction. This representation reveals information about the complete scene that was unavailable to the previously depth-surface based system, supporting the detection of fully- and partially-occluded objects. An over-complete set of cuboid candidates is proposed from the occupancy map and the optimisation applied to find the best cuboid set.

3.2.1 Cuboid Candidates

Potential cuboid candidates are proposed from the volumetric occupancy map to provide an over-complete set to the object detection system. The proposed cuboids are aligned to the Manhattan frame, and represent scene regions that may be an object.

Surface edges of the occupancy map are found by scanning binary gradients in each scene direction. Surface points are clustered in each direction and planar fitting is performed on each group using a RANSAC algorithm [20]. Only planes which align with the Manhattan world axis are considered, with angled planes approximated by the closest axis aligned plane. This provides a coarse Manhattan world planar approximation of the scene.

To produce cuboid candidates, each plane is extended by $e = 20\%$ in each direction to compensate for errors in locating the planar region boundaries during plane fitting. Intersections between extended planes are found; when two planes intersect, their non-extended

outer bounds are taken as an object cuboid candidate [14]. Additional candidates are created from detected horizontal planes, often representing the top supporting surface of scene objects, extended to the floor.

3.2.2 Optimisation

The goal of the optimisation is to select a set of object cuboids that best represent the volumetric occupancy map and the human motion maps without redundancy (i.e. overlapping cuboids). The function, constructed to overcome depth occlusion with terms designed to account for the complete scene, consists of three components: a unary potential U for each candidate, incorporating cuboid solidness, a penalty for protrusion into free space and a cost reduction for support surfaces detected from observed sitting actions; a pairwise overlap term O, ensuring selected cuboids do not overlap each other in the scene; and a coverage term C that accounts for occluded and visible scene regions. The minimisation is defined as:

$$\min_{x}\{U(x)+O(x)-C(x)\}, \tag{2}$$

where x is a set of object cuboids. The cost function optimises the set of cuboids to represent the complete scene, favouring cuboid sets with low individual cuboid costs whilst ensuring cuboids have small overlap with other cuboids, cover much of the occupied space, and do not exist in free space.

Unary potential. The unary term defines the individual cost for each cuboid candidate, x_i, and consists of three terms:

$$U(x_i) = Solidness(x_i) + FreeSpace(x_i) - Support(x_i). \tag{3}$$

Solidness accounts for the 3D occupancy of the candidate and is defined as $1-({}^{v_o}/_{v_a})$, where v_o is the number of occupied voxels within the cuboid and v_a is the total number of voxels within the cuboid. Solidness is weighted by constant μ_1.

FreeSpace is a penalty applied when the cuboid extends into space determined as unoccupied by the support surface mask, B_{stand}. This applies the constraint that scene geometry cannot exist in space recognised as empty by human motion. B_{stand} is segmented into 100 superpixels L_{stand}; each L represents an area of occupied ($=1$) or free space ($=0$) in the scene. Free space superpixels within the overhead projection of a cuboid candidate x_i contribute to the penalty through $FreeSpace = \mu_2 \sum_{L_{stand}(x_i)}(1-L)$.

Support reduces unary cost when a cuboid encloses regions of recognised sitting poses. This applies the constraint that the action of sitting must require a support surface which is likely to be a scene object. B_{sit} is segmented into 100 superpixels L_{sit}, each L representing recognised sittable surfaces ($=1$) or otherwise ($=0$). Superpixels covered by cuboid candidate x_i contribute to the unary cost by $Support = \mu_3 \sum_{L_{sit}(x_i)} L$.

Pairwise overlap. Cuboid candidates often overlap with other candidates that are generated from similar planes. The overlap term ensures that cuboids selected in the optimisation do not heavily intersect each other; overlapping cuboid candidates represent objects covering the same scene space. The pairwise overlap between two cuboids x_i and x_j is defined as their 3D intersection over union (IoU), denoted o_{ij}. Cuboid sets containing two cuboids with $o_{ij} \geq 0.3$ are not permitted in the optimisation. When $o_{ij} < 0.3$, pairwise overlap becomes $O(x) = \lambda \sum_{\{i,j\}\ \forall\{0<o_{ij}<0.3\}} o_{ij} x_i x_j$.

Coverage term. The coverage term ensures minimisation must attempt to fit cuboids to the scene; without it, the optimal solution would select no cuboids. The term determines how much of the complete scene is covered by a cuboid candidate x_i using the volumetric occupancy map V and free space mask B_{stand}. The area a_L of each superpixel L_{stand} is calculated so the area covered by any x_i becomes $A(x_i) = \sum_{L_{stand}(x_i)} L \times a_L$. Volumetric coverage is defined as $G(x_i) = {}^{v_o}/_{n_V}$ where v_o is the number of occupied voxels within cuboid x_i and n_V is the total number of voxels in the scene. For any given x, the coverage term becomes $C(x) = \gamma \sum_{x_i}(A(x_i) + G(x_i))$.

Solution. The optimisation is solved via branch and bound [11]. Term weights are set to $\mu_1 = 0.5$, $\mu_2 = 0.5$, $\mu_3 = 0.05$, $\lambda = 0.1$ and $\gamma = 2$ through experiment.

4 Evaluation

In this section we assess the proposed system to evaluate the volumetric occupancy and object detection performance. The accuracy of free space estimation is determined by voxel-level precision, recall and IoU and compared to previous approaches using synthesised motions maps. Object detection performance is compared to baselines on new RGBD captures of dynamic scenes. Existing datasets for evaluation of object detection are typically limited to single RGBD images or sequences from a moving RGBD sensor, and critically do not include long-term human motion. Therefore, we introduce a new benchmark dataset for evaluation of single-view RGBD combined with human motion. This dataset will be made available to support further research.

4.1 Quantitative Evaluation

Volumetric Occupancy

The proposed system relies on long-term sequences of in-use indoor scenes, so that human motion can be applied effectively. To our knowledge, no current RGBD datasets focus on this problem, delivering sequences of static and dynamic scenes from mobile sensors rather than long-term static sensor capture of dynamic scenes. Previous approaches attempting to infer the complete scene using single-shot captures [4, 25] cannot be applied to our approach; this creates difficulty in evaluation as we cannot readily evaluate on existing datasets.

To overcome this, we synthesise maps of human motion from ground truth annotations of the NYU dataset [17]. The ground truths provided by [4] supply 3D annotations of scene planes, boxes and object models that we voxelise to create ground truth volumetric occupancy maps of each scene. To synthesise the motion maps, the ground truth occupancy map is projected into the overhead view. Regions that permit standing actions are estimated as voxel columns where no object annotations exist, to create B_{stand}. Surfaces that support sitting actions are estimated from included support surface data on functionably sittable objects, to produce B_{sit}. M_{stand} and M_{sit} are estimated as B_{stand} and B_{sit} and eroded to replicate real-world inconsistencies. During testing, the volumetric occupancy map $V_{i,j,k}$ is initialised on depth rendered from the ground truth annotated models, as with [25] and [4], to overcome annotation error during evaluation.

The metric used for evaluation is the intersection over union of voxels predicting free or occupied space. Only voxels occluded by the depth surface are considered in the evaluation. As a baseline we use the IoU of the non-carved occupancy map (*Base-NC-completion*), and

Method	Prec.	Recall	IoU
Firman *et al.* [4]	66.5	69.7	50.8
Song *et al.* [25]	66.3	96.9	64.8
Base-NC-completion	56.9	**99.8**	56.8
Ours-completion	**67.1**	**99.8**	**67.0**

Table 1: **Voxel-level scene completion on rendered NYU annotations**

Method	Prec.	Recall
Base-D-detection	73.8	52.6
Base-NC-detection	80.1	50.0
Ours-detection	**88.1**	**80.2**

Table 2: **3D object detection on our human-scene dataset**

compare our volumetric carving (*Ours-completion*) to the approaches of Song *et al.* [25] and Firman *et al.* [4].

In Table 1, we compare our results to Firman *et al.* [4] and Song *et al.* [25], who predict the complete scene from a single depth map. Through trained models, their approaches contribute precise occupancy predictions in regions surrounding scene objects; from the observation of human motion, we achieve enhanced performance through the resolution of scene-level occlusions that our approach determines. Initiating the occluded regions to occupied space to be carved demonstrates high voxel-level recall, as we are removing known free space over predicting unknown regions. This concept reveals the full extent of occluded regions without geometric modelling, demonstrating that human-centric scene analysis can augment current scene understanding approaches to enhance complete reconstruction. Song *et al.* [25] show additional results employing object recognition in *semantic* scene completion; we hope to incorporate this concept in future work.

Object Detection

To test object detection, we captured a number of RGBD scenes with the Kinect v2. The captures initialise with a static scene, before a person explores the space to provide joint positions. The captures run between 1-2 minutes with a single person exploring the complete scene. We assume that "real-life" captures would explore the complete space through continued use; our shortened sequences make capture more accessible. The captured scenes are designed to contain complex but realistic occlusions that could occur in cluttered spaces. Bounding box annotations of large furniture objects (chairs, tables, cabinets, shelves, etc.) are produced through manual labelling for object detection evaluation.

The average precision and recall of detected objects over all scenes is used for evaluation. An object is considered detected if its IoU is greater than 0.3 with a predicted bounding box, to account for 3D detection sensitivity. To demonstrate how human motion can improve object detection, we compare our approach to a baseline system with all optimisation terms incorporating human motion removed. We compare the full system applied to the carved occupancy map (*Ours-detection*) to the baseline system applied on only visible depth (*Base-D-detection*) and on the non-carved occupancy map (*Base-NC-detection*).

The results in Table 2 show how object detection performance improves when volumetric occupancy is utilised over depth. The volumetric representation stimulates the detection of partially and fully occluded objects, providing a significant increase in object recall that is unattainable in current single-view depth systems. Non-occluded object detection additionally benefits from the representation, as the full extent of scene objects is revealed by human

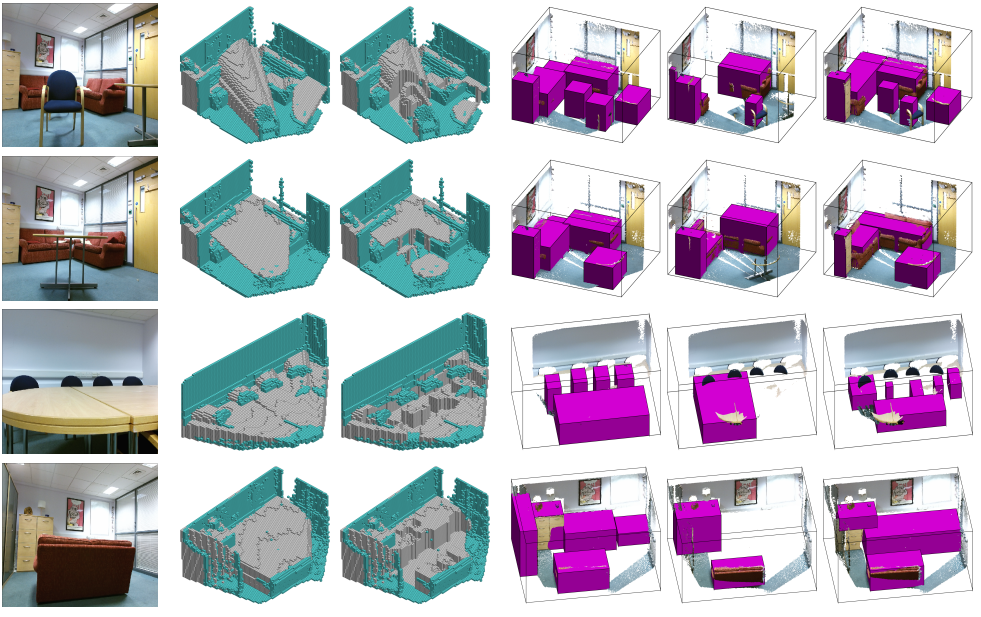


Figure 3: **Qualitative results.**

motion; when the scene is not carved, the process must fit to visible surfaces only.

4.2 Qualitative Evaluation

Examples of volumetric occupancy map carving and the object detection output are shown on our captured data in Figure 3. Comparison between non-carved (*Base-NC-completion*) and carved (*Ours-completion*) volumetric occupancy maps illustrates how occluded scene regions are reconstructed by application of human motion. In exposing these occluded regions, unobserved objects can be detected to enhance structured reconstruction of the complete scene (shown in *Base-D-detection* vs. *Ours-detection*).

Significant contributions are shown in the first row as a fully occluded chair is reconstructed, and the second row where the full extent of the self-occluding table is detected. The third row demonstrates a system limitation: when the full scene is not explored by a human, false positive object detections can arise as cuboid candidates are incorrectly fit to unexposed free space.

5 Conclusions and Future Work

In this paper we presented analysis of human motion as a solution to occlusions present in single-view depth. We introduced a human-centric scene representation to model human movement, which is utilised to reveal occluded scene regions in a volumetric occupancy map when applied with body occupancy constraints. An object detection system designed to exploit the revealed occupancy allows detection of visible and occluded objects from single-view depth sequences. We have shown that this concept enables the complete reconstruction

of fully-, partially- and self-occluded objects to their full extent. We evaluated against two state-of-the-art single-view depth-based scene completion systems to demonstrate the relevance of our procedure. As the first approach applying human motion towards occlusion reconstruction, we provide a benchmark dataset with baselines indicating the contribution human motion analysis can provide to complete scene reconstruction.

In line with single-shot image reconstruction approaches, the current implementation is limited to modelling static scenes as interactions that modify scene structure (such as a person moving a chair) are ignored. Recognising object displacement will be considered by monitoring semantically labelled object instances over long-term non-simulated dynamic scene captures. Further work will investigate this limitation by employing an improved model of human actions and motion.

Acknowledgements

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership. Details about the data underlying this work, along with the terms for data access, are available from: http://dx.doi.org/10.15126/surreydata.00812228. We thank Teo DeCampos for his assistance in capturing data and continued support.

References

[1] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *CVPR*, 2015.

[2] Vincent Delaitre, David F Fouhey, Ivan Laptev, Josef Sivic, Abhinav Gupta, and Alexei A Efros. Scene semantics from long-term observation of people. In *ECCV*. Springer, 2012.

[3] Felix Endres, Jürgen Hess, Nikolas Engelhard, Jürgen Sturm, Daniel Cremers, and Wolfram Burgard. An evaluation of the rgb-d slam system. In *ICRA*, 2012.

[4] Michael Firman, Oisin Mac Aodha, Simon Julier, and Gabriel J Brostow. Structured prediction of unobserved voxels from a single depth image. In *CVPR*, 2016.

[5] David F. Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A. Efros, Ivan Laptev, and Josef Sivic. People watching: Human actions as a cue for single-view geometry. In *ECCV*, 2012.

[6] Andreas Geiger and Chaohui Wang. Joint 3d object and layout inference from a single rgb-d image. In *GCPR*, 2015.

[7] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? In *CVPR*, 2011.

[8] Ruiqi Guo and Derek Hoiem. Support surface prediction in indoor scenes. In *ICCV*, 2013.

[9] Ruiqi Guo, Chuhang Zou, and Derek Hoiem. Predicting complete 3d models of indoor scenes. *arXiv preprint arXiv:1504.02437*, 2015.

[10] Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. From 3d scene geometry to human workspace. In *CVPR*, 2011.

[11] Hao Jiang and Jianxiong Xiao. A linear approach to matching cuboids in rgbd images. In *CVPR*, 2013.

[12] Yun Jiang, Marcus Lim, and Ashutosh Saxena. Learning object arrangements in 3d scenes using human context. In *ICML*, 2012.

[13] Yun Jiang, Hema Koppula, and Ashutosh Saxena. Hallucinated humans as the hidden context for labeling 3d scenes. In *CVPR*, 2013.

[14] Hansung Kim and Adrian Hilton. Block world reconstruction from spherical stereo image pairs. *CVIU*, 2015.

[15] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 2013.

[16] Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *ICCV*, 2013.

[17] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[18] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, 2011.

[19] Zhile Ren and Erik B Sudderth. Three-dimensional object detection and layout prediction using clouds of oriented gradients. In *CVPR*, 2016.

[20] Ruwen Schnabel, Roland Wahl, and Reinhard Klein. Efficient ransac for point-cloud shape detection. In *Computer Graphics Forum*, 2007.

[21] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *CVPR*, 2013.

[22] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *ECCV*, 2012.

[23] Nathan Silberman, Lior Shapira, Ran Gal, and Pushmeet Kohli. A contour completion model for augmenting surface reconstructions. In *ECCV*. Springer, 2014.

[24] Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *CVPR*, 2016.

[25] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *CVPR*, 2017.

[26] Yixin Zhu, Chenfanfu Jiang, Yibiao Zhao, Demetri Terzopoulos, and Song-Chun Zhu. Inferring forces and learning human utilities from videos. In *CVPR*, 2016.