Human-Centric Scene Understanding from Single View 360 Video

Sam Fowler Hansung Kim Adrian Hilton
Centre for Vision, Speech and Signal Processing,
University of Surrey, Guildford, UK
{sam.fowler, h.kim, a.hilton}@surrey.ac.uk
http://cvssp.org/projects/s3a/AffordRecon

Abstract

In this paper, we propose an approach to indoor scene understanding from observation of people in single view spherical video. As input, our approach takes a centrally located spherical video capture of an indoor scene, estimating the 3D localisation of human actions performed throughout the long term capture. The central contribution of this work is a deep convolutional encoder-decoder network trained on a synthetic dataset to reconstruct regions of affordance from captured human activity. The predicted affordance segmentation is then applied to compose a reconstruction of the complete 3D scene, integrating the affordance segmentation into 3D space. The mapping learnt between human activity and affordance segmentation demonstrates that omnidirectional observation of human activity can be applied to scene understanding tasks such as 3D reconstruction. We show that our approach using only observation of people performs well against previous approaches, allowing reconstruction of occluded regions and labelling of scene affordances.

1. Introduction

As smart technology and connected sensors become more prominent in everyday life, users are becoming naturally accustomed to smart environments. The full potential of different sensors towards smart home applications has not yet been realised. Spherical cameras allow image and video capture across a full 360 degree view: whilst commonly used for both casual photography and 360 degree film making, they also present interesting research avenues within indoor scene understanding. As recent advances reduce the size and cost of such sensors, their importance in understanding complete indoor scenes, such as in home environments, increases.

This paper investigates the application of a spherical sensor located centrally within an indoor room to analyse human behaviour towards solving scene understanding tasks. The approach utilises observation of human motion and actions for semantic scene understanding. Exploiting the observation of human activity within the scene helps to recognise scene affordance, and overcomes single image understanding limitations such as object occlusions. Omnidirectional spherical capture of human activity has the advantage of localising people and the actions they perform throughout the complete scene, providing unique cues into the structure and affordance of the space which are unobtainable through visual scene analysis.

General scene understanding solutions largely utilise standard perspective or depth cameras to achieve semantic segmentation, object detection and scene reconstruction tasks. All techniques involving these sensors are inherently restricted to a limited field of view, losing valuable information regarding the complete scene the sensor is within. Systems utilising multisensor rigs [4], requiring user interaction [20] or by direct inference [26], implement attempts to overcome this limitation but each suffer their own disadvantages, such as high implementation costs. In recent years, panoramic images have been applied to scene understanding tasks to contribute the information necessary to reconstruct the complete scene from a single location. Recent low-cost compact sensors [1] now present the opportunity to easily access spherical video, allowing investigation into human activity over the complete scene.

Given human motion within a scene, the proposed system estimates human poses over long term spherical video captures to record complete scene human motion and actions. A probabilistic map recording localised human actions is generated, exploiting the geometric nature of the spherical sensor to estimate the relative 3D position of users over time. A synthetically created dataset of such activity maps is produced and utilised to train a semantic segmen-

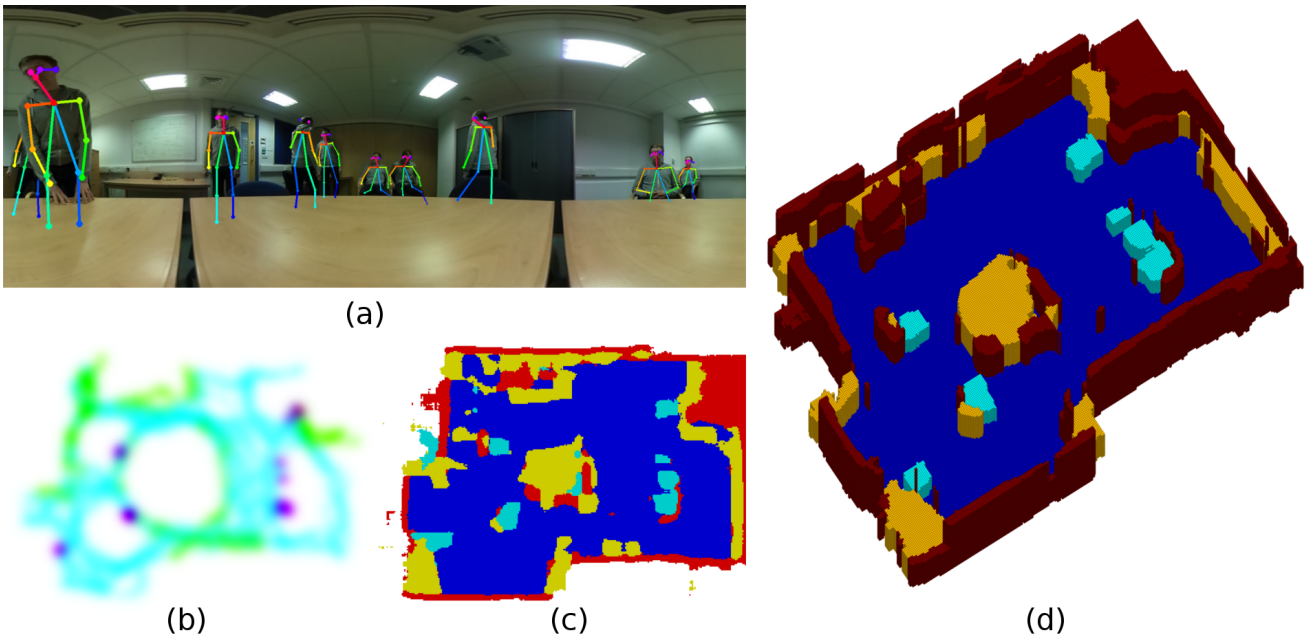


Figure 1. **Proposed system**: Human poses are estimated over long-term spherical captures (a) to determine human activity over the complete scene. Actions performed over time are recorded in a top-down map of scene activity (b) (blue=walking, purple=sitting, green=using), applying the geometric nature of the spherical sensor to estimate the 3D position of poses within the scene. A trained model predicts a scene affordance segmentation (c) from the activity map, localising regions of affordance throughout the complete scene (blue=walkable, teal=sittable, yellow=usable, red=structure). A complete scene reconstruction (d) is composed from the activity and affordance maps, contributing 3D scene structure and affordance estimations from human activity alone.

tation network to predict affordance labels and scene structure from a top-down 2D perspective. A 3D reconstruction is composed from the affordance prediction and captured human activity, generating a volumetric affordance segmentation of the complete 3D scene.

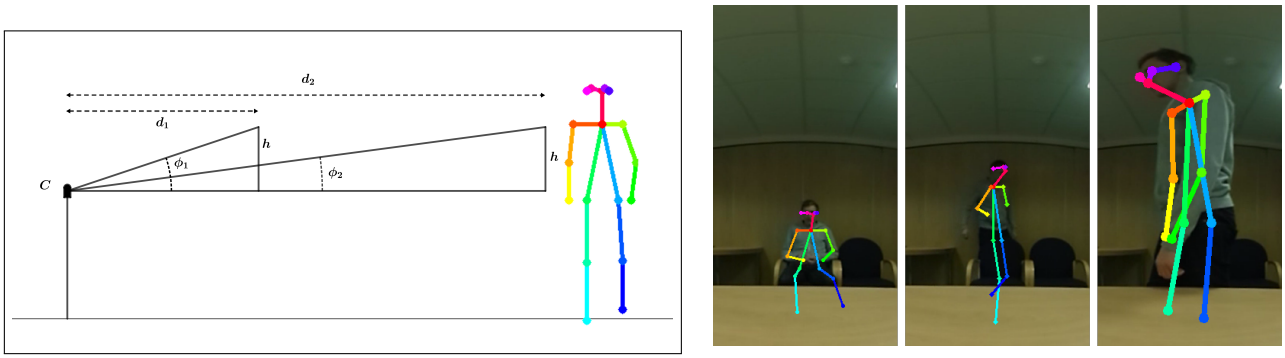
The key contributions of this work are: (1) A human-centric approach to scene understanding from single view spherical video to estimate occluded scene structure and affordance from motion and behaviour of people; (2) A deep fully convolutional encoder-decoder network to map observation of human pose and action to scene object affordance labels for the complete scene; (3) An affordance segmented 3D reconstruction of a complete scene utilising only captured human activity, providing unique scene cues that overcome limitations such as scene occlusions.

2. Related Work

We study indoor scene understanding works investigating 3D scene reconstruction and affordance prediction, with particular interest in the application of spherical imagery and human motion analysis towards these tasks.

Scene reconstruction has been a topic of much interest in the computer vision community for many years. Continuing research investigates perspective monocular [25], stereo [9] and depth [18] based approaches to produce geometric and structured 3D scene reconstructions. An increasing volume of research explores 3D reconstruction from panoramic and spherical imagery, exploiting the wider field of view to reconstruct complete indoor scenes [14, 15, 30, 31, 32, 33]. Yang and Zhang [32] recover 3D room layout from indoor panoramas, performing reconstruction on a constraint graph incorporating the relationships between scene lines and superpixels. Zhang et al. [33] generate 3D object hypotheses from geometric and semantic contextual constraints, sampling the hypotheses to create a 3D scene model. Kim et al. [15] utilise spherical sensors in a stereo pair to estimate scene depth, fitting planes and cuboids to construct a 3D room and object layout estimation, extended to 3D structure estimation with learnt object material properties [14].

Affordance prediction covers two major forms: scene level and object level affordances. Object level predictions localise interactable object features such as “graspable” or “container”, and largely concerns human-robot interaction applications. Techniques towards this goal have been implemented utilising image-based object detection frameworks [6, 21] and through analysing human- and robot-object interactions [16, 17, 35]. Of interest to this work, scene level predictions consider how structural scene objects, such as “sittable” and “walkable” supporting surfaces, afford hu-



(a) (b)

Figure 2. (a) The relative distance (e.g. d_1, d_2) of a detected person from the sensor is geometrically estimated using the neck joint (the height of the camera and human subject are unknown). (b) Examples of different poses at similar azimuths ϕ, with varying elevations θ dependent on action (left: sitting) and distance (middle: far, right: near).

man activity, characterising how a human can interact with the scene. Many of these approaches manipulate simulated human poses within a scene to achieve affordance detection [10, 11, 13, 22, 34], with few adapting semantic segmentation techniques to provide pixel-wise affordance labels [24]. Gupta et al. [11] predict a human "workspace" from single indoor scene images, defining allowable scene actions afforded by simulated human poses. Jiang et al. [13] consider human-object context, hallucinating humans to display hidden context when labelling 3D scenes. Zhu et al. [34] recognise the concept of human utilities, inferring forces acting between simulated poses and scene objects to predict scene affordances.

Surprisingly, analysis of non-simulated human activity towards both indoor scene reconstruction and scene-level affordance labelling in monocular settings is notably scarce. Real-person analysis has been applied outside, in applications such as surveillance monitoring: Rother et al. [23] demonstrate deriving 3D information from tracked people to determine scene geometry whilst Ballen et al. [3] learn navigation paths to predict future activities and semantically segment scenes. Indoors, Fowler et al. [8] expand Fouhey et al.'s [7] concept of providing unique cues from activity analysis to improve standard 3D reconstruction approaches, whilst Delaitre et al. [5] study semantic object recognition from long-term observation of human-object interactions.

We hope that recent large-scale pose datasets [28] and advances in pose estimation accuracy [29] will motivate further research applying real human activity analysis to indoor scene understanding tasks. Towards this goal, we will make the simulated and captured datasets introduced in this work available to the research community.

3. Complete Scene Understanding

This section details our approach to producing an affordance labelled 3D scene reconstruction by considering human activity in the complete scene, as seen in Figure 1. Section 3.1 discusses how human activity is recorded from spherical imagery, describing how the relative depth of a person is estimated. The training of an encoder-decoder network for mapping the human activity representation to scene affordance labels is described in Section 3.2. Section 3.3 then outlines how a labelled volumetric 3D reconstruction of the scene is formed from the human activity and affordance labelling results.

3.1. Human Monitoring

Human movement and actions occurring within a scene are captured in a probabilistic 2D map from a top-down scene perspective. Such a representation records the co-occurrence and spatial localisation of actions over time. To achieve this, the position of people visible within the spherical video capture of the complete scene must be estimated in a 3D space. In general, the absolute depth of an object is unavailable with no known reference lengths; this work estimates the 3D position of a person within the scene from their relative distance to the sensor. An activity map M_{act} is created to record the relative positions of actions occurring over time, represented as a $300 \times 300 \times 3$ sized normalised probabilistic array throughout this work.

3.1.1 Human Depth Estimation

Human poses are estimated by applying Convolution Pose Machines (CPMs) [29] to each equirectangular frame (representing $360°$ horizontally and $180°$ vertically) of the spherical video stream. Each pose prediction consists of a 2D equirectangular image coordinate for each joint in the skeleton body model. Over a long capture, this produces a collection of estimated locations for human poses within the 2D equirectangular scene.

To map human actions onto M_{act}, the 3D position of a

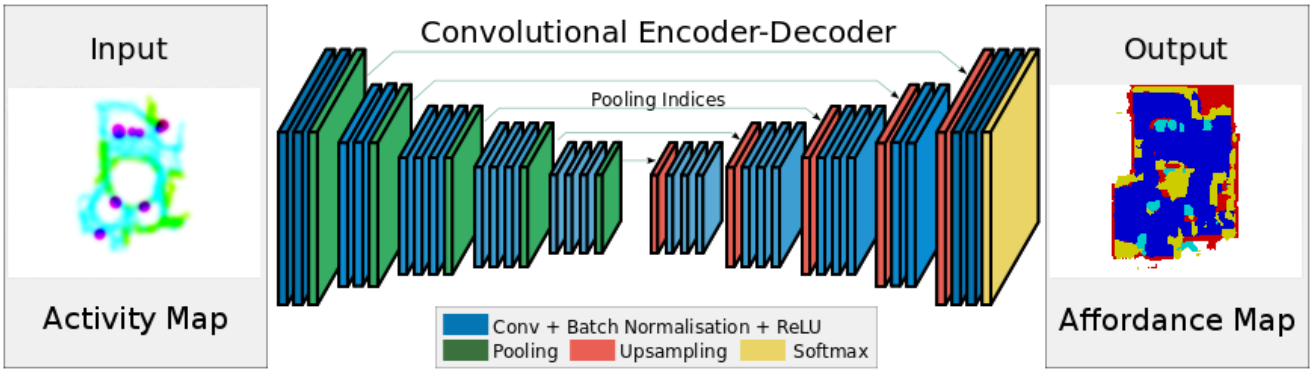


Figure 3. The deep fully convolutional SegNet [2] architecture is trained from scratch to segment regions of human affordance from captured activity maps. Image adapted from [2].

predicted pose relative to the camera is calculated. Through experimentation, only the predicted neck joint is considered in the calculation; it is often unoccluded in a variety of scenes and its location is largely invariant to body motions such as rotation. The spherical sensor allows the predicted neck joint coordinate to be easily projected into spherical coordinates $[\theta, \phi, r]$, where θ is the azimuth angle, ϕ is the elevation angle and r is the radial distance [14]. As the azimuth θ provides the polar angle from the camera to the joint in a top-down perspective, only the radial distance of the body from the camera, d, needs to be calculated.

Before pose estimation, the camera is aligned to the scene with [33] to ensure the plane at $\phi = 0$ is parallel to the ground plane. Now, when standing anywhere in the scene, the height h of a persons neck joint above or below the horizontal $\phi = 0$ plane remains constant throughout a capture, as shown in Figure 2. In this configuration, a large elevation angle ϕ represents a person close to the camera, with a lower elevation representing that person further away. As such, h can be set as a reference length (i.e. $h = 1$) to simplify the geometric relationship $d = h/\tan\phi$.

As a person moves away from the sensor, a quantisation noise is introduced to the estimated depth due to a reducing number of pixels representing the same elevation loss. To suppress this, the sequential list of estimated depths d is smoothed to create a continuous path of human activity.

3.1.2 Activity Maps

Each estimated 2D pose now has an estimated polar coordinate $[d, \theta]$. However, as the estimated d is dependent upon the joint elevation, any action such as sitting down will present as an increase in depth due to the lowering joint elevation. A simple model is trained to recognise actions from each predicted pose, considering the actions "walking"/"standing", "sitting" and "using"; whilst ignoring fine-grained actions, these coarse labels recognise the key affordance types that relate to scene structure (i.e. walkable, sittable and usable surfaces).

To project human activity onto M_{act}, the estimated polar coordinate of each predicted "standing" pose is converted to Cartesian ground plane coordinates by $[x,\ y] = [d\cos(\theta),\ \ d\sin(\theta)]$. Poses recognised as "sitting" are located at the last recorded "standing" location, as the person moves from standing to sitting in the same location. The location of "usable" actions is dependent upon the sitting or standing action that is occurring concurrently, and is placed at the last location of the relevant action. A Gaussian distribution is applied at each action location to represent regions of likely occupancy, and the outcome is recorded in channels of M_{act} for each action type.

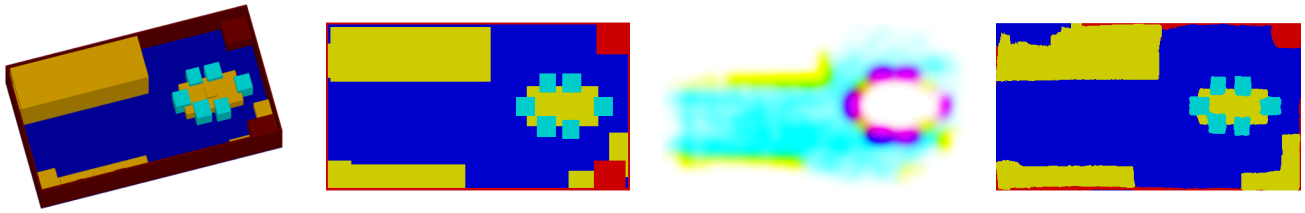
M_{act} now provides a map of the scene in terms of human motion and actions, localising areas of walkable, sittable and usable scene regions as seen in Figure 1(b).

3.2. Human Affordance Labelling

A deep fully convolutional encoder-decoder network is trained to segment regions of affordance and scene structure from the human activity maps described in Section 3.1. This process allows detected human poses to be mapped to affordance regions, providing details of scene structure and of semantic object categories. The network, adapted from SegNet [2], is trained from scratch on synthetic activity and affordance maps to predict a top-down affordance segmentation M_{aff} from an activity map M_{act}, as shown in Figure 3.

The SUNCG dataset [27], designed for scene understanding tasks from rendered scene images, is adapted to train the network. SUNCG contains over 45000 3D CAD models of indoor scenes complete with semantically annotated 3D object models. These models are processed to supply top-down room layouts with affordance labelled scene objects and structure, simulating the desired prediction M_{aff}.

Each 3D scene model is voxelised to provide a volumetric occupancy representation of the scene. The voxel scenes



(a) (b) (c) (d)

Figure 4. (a) Voxelised ground truth scene reconstruction from SUNCG dataset [27]. (b) Ground truth affordance segmentation derived from the voxelised scene. (c) Human activity about the scene simulated from the ground truth affordance map. (d) Affordance segmentation prediction, M_{aff}, from applying the trained model to only the simulated human activity map (c).

are projected into 2D top-down perspectives, producing 2D plans of the scene layout. Each vertical column of scene voxels now represents a single ground truth image pixel of M_{aff}; the mode semantic object label over each column is taken as the object label for that pixel, providing a semantically segmented scene plan.

The semantic object labels are grouped into 5 object-affordance classes: unlabelled (regions outside of the scene), structure (walls, miscellaneous objects), sittable (e.g. chairs, sofas), walkable (i.e. floor) and usable (e.g. tables, cabinets). By mapping semantic object labels to affordance labels, a large collection of scene maps segmented with affordance regions is created, to be applied as the ground truth or expected outcome of the network.

Synthetic activity maps are created from the ground truth affordance maps to produce the training and test datasets. Motion paths through the scene are simulated by randomly selecting consecutive points in "walkable" scene regions, as sittable and usable action regions are estimated from these in M_{act}. The A* search algorithm [12] is applied to find a path between each pair of consecutive points that does not pass through any of the structure, sittable or usable scene regions to produce a walking motion path throughout the scene.

Sittable object regions in the ground truth affordance maps are dilated to create an overlap with the simulated motion path; a random number of the overlapped points are selected to represent sitting locations around the sittable objects, simulating how the real activity maps are created. A similar process creates locations for actions around usable objects, with a larger random selection of points selected around these objects to represent the entirety of the usable surface.

Motion path points selected for each of the three activity types are placed into separate image channels. Gaussian distributions, representing the size and occupancy likelihood of the simulated person, are applied at each motion path point to produce the synthetic activity maps, M_{act}, from ground truth affordance maps, M_{aff}. The dataset generation process is presented in Figure 4.

The ground truth affordance maps and motion maps are augmented through flipping and rotating, and resized to the 360×480 network input size. A training and test set of 160000 and 8000 of the synthetic image pairs are selected respectively, to be released to the public with the trained model. The model, trained for 27000 epochs, is capable of predicting a scene affordance segmentation from human activity within the scene.

3.3. 3D Reconstruction

A 3D representation of the scene is reconstructed from the long term human activity and predicted affordance segmentation. This produces a 3D voxel grid representation of scene occupancy to reveal structure whilst classifying 3D affordance regions with the predicted affordance segmentation.

The reconstruction technique applies scene structure constraints on the predicted affordance segmentation, alongside the assumption discussed in [8] that space occupied by a human body must be free space. The reconstruction is produced solely by applying these constraints to M_{act} and M_{aff} to estimate a 3D voxel grid, with no current consideration of visual scene information.

A number of binary maps are constructed from M_{act} and M_{aff} for each of the relevant affordances and structures: B_{walk}, B_{sit}, B_{use}, B_{walls} and $B_{structure}$. M_{act} is thresholded (> 0.1 through experimentation) to provide areas of likely human occupancy for each of the three actions. The combination of the walking regions from this with the walkable surfaces predicted in M_{aff} creates the map of walkable regions, B_{walk}. To determine the scene walls, the outer region of B_{walk} is masked by the structure labels predicted in M_{aff} to produce B_{walls}. The remaining regions of predicted structure create $B_{structure}$, that represents scene objects and structure with no recognised affordance.

When reconstructing synthesised scenes, predicted sittable affordance regions are taken directly as the sittable region map B_{sit}. However, the smoothing required in the

creation of real-scene activity maps introduces spatial artefacts into M_{act} that cause spotty noise in the predicted sitting affordances. Therefore, when considering real scenes, the sitting regions from M_{act} are applied to mask the predicted sittable regions to create B_{sit}. Finally, the predicted usable affordance regions are masked by both B_{walk} and B_{sit} to create B_{use}.

With no known absolute reference lengths, the absolute size of the room cannot be estimated. As such, the 3D reconstruction map $V_{i,j,k}$ is initialised as a $i = 300 \times j = 300 \times k = 100$ voxel grid to fit the scene within. Each voxel $V_{i,j,k}$ is initialised as empty $(= 0)$ so that scene structure can be formed from the affordance maps as follows:

Algorithm 1 3D Reconstruction Algorithm

for $layerHeight = 1 : k$ **do**
  $layer = V(:, :, layerHeight)$
  $layer(B_{wall}) = L_{wall}$ ▷ Set walls
  if $layerHeight == 1$ **then**
    $layer(B_{walk}) = L_{walk}$ ▷ Set floor
  if $layerHeight <= \frac{k}{3}$ **then**
    $layer(B_{sit}) = L_{sit}$ ▷ Set sittable
  if $layerHeight <= \frac{k}{2}$ **then**
    $layer(B_{structure}) = L_{structure}$ ▷ Set structure
    $layer(B_{use}) = L_{use}$ ▷ Set usable
  $V(:, :, layerHeight) = layer$

where the integer labels of V for each affordance class are represented by $L_{walk} = 1$, $L_{sit} = 2$, $L_{use} = 3$, $L_{structure} = 4$, $L_{wall} = 5$ and k represents the maximum height of the scene.

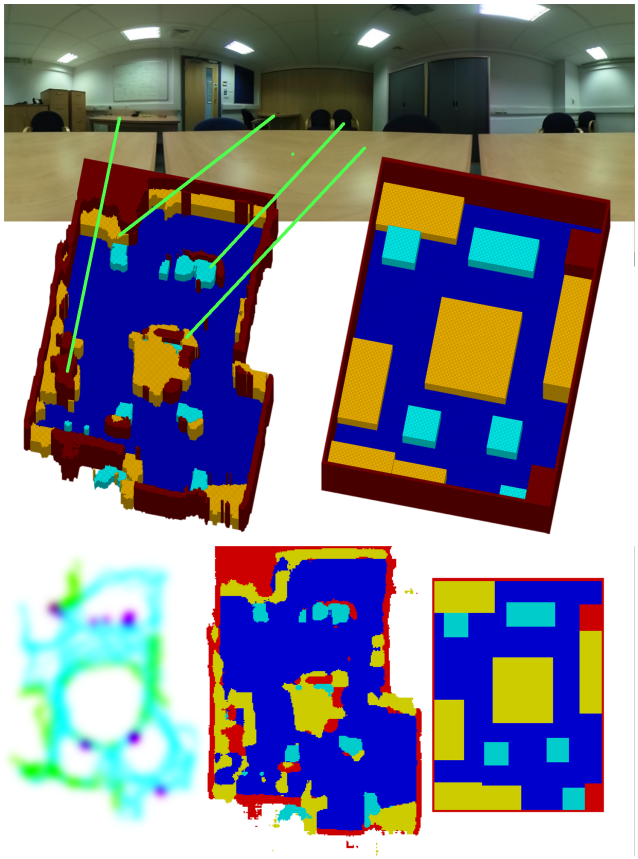


Figure 5. 3D reconstruction from predicted affordance labels (middle left) with ground truth scene model (middle right). Activity map (bottom left) with predicted (bottom middle) and ground truth (bottom right) affordance segmentations. Affordance labels: blue=walkable, teal=sittable, yellow=usable, red=structure.

4. Evaluation

In this section, we assess the performance of the proposed complete scene 3D reconstruction and affordance labelling components on real and synthetic scenes.

As far as we are aware, no current approaches consider human activity within spherical video captures towards the goal of reconstruction or affordance segmentation. Relevant algorithms considering complete scene reconstruction from spherical imagery rely on single panoramic images of the scene, resulting in no comparison datasets for evaluation. A small number of spherical timelapse video captures exist online in the form of casual photography, but these do not incorporate any form of 3D ground truth annotation. To this end, we introduce a dataset of long-term spherical captures containing human activity within real indoor scenes, complete with measured ground truth annotations. The captured scenes were monitored for 20 minutes with a human actor moving into available space and utilising "sittable" and "usable" surfaces to provide data for the scene. This approach was taken to simulate the potential of longer captures of indoor scenes as may be available in future applications. The dimensions of the scenes and the size and position of objects within them were manually measured as a ground truth 3D reconstruction. Each annotated object model was manually classified with affordance labels to evaluate the affordance prediction model.

Also available is the synthetic dataset created as part of the affordance prediction model training process. Derived from annotated 3D CAD models, this dataset provides ground truth 3D reconstructions and affordance maps for the composed activity maps. The training set within this dataset allows us to assess the efficiency of the proposed affordance segmentation and reconstruction techniques.

Quantitative results for reconstruction and affordance segmentation are discussed in Sections 4.1 and 4.2. Qualitative results are presented in Figures 5 and 6.

Method	Prec.	Recall	IoU
Motion [8]	0.231	0.243	**0.859**
Ours-GT-Aff	**0.713**	0.894	0.783
Ours-GT-Aff+Act	0.669	**0.895**	0.730
Ours-Pred-Aff	0.430	0.759	0.506
Ours-Pred-Aff+Act	0.427	0.760	0.501

Table 1. **3D reconstruction on the synthetic dataset**

Method	Prec.	Recall	IoU	Obj. det.	Aff. det.
Motion [8]	0.375	0.329	0.723	0.911	-
Ours-GT-Aff	0.799	**0.904**	**0.726**	**1**	**1**
Ours-GT-Aff+Act	**0.831**	0.747	0.644	0.986	0.986
Ours-Pred-Aff	0.457	0.486	0.305	0.811	0.556
Ours-Pred-Aff+Act	0.477	0.475	0.311	0.797	0.556

Table 2. **3D reconstruction on captured scenes**

4.1. Complete 3D Reconstruction

The proposed reconstruction approach is evaluated on the synthetic and real datasets to validate the approach against the state of the art [8], assessing how the introduction of affordance prediction improves reconstruction over applying only human motion. Four variants of the proposed method are evaluated against [8] in an ablation study: our technique applying only the ground truth affordance map (*Ours-GT-Aff*) and applying it with the activity map (*Ours-GT-Aff+Act*), and our technique applying only the predicted affordance map (*Ours-Pred-Aff*) and applying it with the activity map (*Ours-Pred-Aff+Act*).

Voxel-level precision, recall and intersection over union (IoU) [27] are applied as the evaluation metric; all reconstructed voxels are treated as one class as [8] has no semantic segmentation component. Only voxels within the ground truth 3D reconstruction are considered in the evaluation.

When evaluating on real scenes, the ground truth reconstruction must be manually aligned with the prediction due to the nature of the relative depth component in the proposed solution. This introduces positional artefacts that will display in voxel-level examination. To accommodate this, we introduce an object detection evaluation that simultaneously accounts for the inaccurate nature of the ground truth box model annotations. Bounding boxes are placed around affordance segmented connected components within the 3D reconstruction and their IoU is compared to objects in the constructed ground truth model. A ground truth object is considered detected if its IoU is greater than 0.3 with the reconstruction to acknowledge 3D detection sensitivities. An affordance wise object detection is further introduced, evaluating how well objects are detected of the correct affordance class. This follows the same IoU validation, but only voxels of the appropriate affordance class are considered.

Discussion

Tables 1 and 2 present 3D reconstruction results of the proposed approaches on the synthetic and real datasets respectively, evaluating against a state of the art approach which utilises only the human activity map in reconstruction. The results show that introducing the scene affordance map to the reconstruction process improves both the precision and recall of occupied voxels. Note that [8] initialises the reconstruction with a fully occupied voxel grid, resulting in comparatively high IoU results.

Results for the ground truth affordance map reconstructions demonstrate that approximations of object height in the reconstruction process do not significantly impact reconstruction accuracy. Whilst determining reconstructed object height solely from predicted affordance labels, the high accuracy of results for *Ours-GT-Aff* and *Ours-GT-Aff+Act* advocate that such an approximation can reconstruct a scene well.

As expected, a performance drop is presented when utilising predicted over ground truth affordance maps due to inaccuracies in the affordance prediction (i.e. *Ours-Pred-Aff+Act* vs. *Ours-GT-Aff+Act*). However, this performance decrease is comparable between the synthetic and real datasets, demonstrating that the synthetically trained model performs reasonably on captured scenes.

Object detection results on the captured dataset demonstrate that 3D scene objects are represented well in the reconstruction using either the ground truth or predicted affordance maps. Affordance-wise object detection performs well, even without including "structure" predictions which are often classified near affordance labelled objects.

Qualitative evaluation shows that interior scene objects such as chairs and tables are generally reconstructed well. Whilst the room layout is inherently estimated, wall placement can inaccurately cut into the scene where exterior scene objects are located. We notice the occasional reconstruction of phantom objects caused by insufficient motion data in certain regions, especially on scene perimeters.

Scene	Pix. Acc.	Mean Acc.	Mean IU	Freq. IU
Syn-aff	**0.967**	**0.953**	**0.889**	**0.943**
Real-aff	0.779	0.529	0.431	0.696
SegNet [2]	0.897	0.874	0.745	0.827

Table 3. **Affordance segmentation results on synthetic and real scenes.**

4.2. Human Affordance Segmentation

The proposed affordance segmentation model is applied to the test set of human activity maps synthesised from the SUNCG 3D scene dataset (*Syn-aff*). To measure model performance, the predicted segmentations are evaluated against the ground truth affordance maps used to generate the activ-

ity maps. The evaluation is performed across all affordance classes and only on pixels within the ground truth scene. Four standard metrics to determine semantic segmentation accuracy are considered: pixel accuracy, mean accuracy, mean IoU and frequency weighted IoU [19].

The same evaluation is performed on the affordance maps predicted from human activity captured in real scenes (*Real-aff*). As with the 3D reconstruction evaluation, for evaluation the ground truth affordance annotations are manually aligned with the predictions due to the relative depth estimation.

As no appropriate solutions are available for direct comparison, we include an affordance segmentation result considering only the scene image and no human motion (*SegNet* [2]). A model trained for semantic segmentation of indoor scenes was applied to cubic images projected from our captured spherical images. The predicted semantic object labels were mapped to affordance labels to provide a baseline for affordance segmentation.

Discussion

Affordance segmentation results are presented in Table 3. Results on the synthetic test set present accuracy exceeding the *SegNet* baseline, demonstrating good performance on mapping human activity to scene affordance labels. A drop in accuracy is seen when applying the model to the real scene dataset, but affordance regions are generally represented well, as visible in Figures 5 and 6. We speculate that fine-tuning the trained network with captured human activity would further improve performance on real scenes.

5. Conclusions and Future Work

In this paper, we presented an approach to reconstruct a complete 3D, affordance labelled representation of an indoor scene, from a single spherical camera using captured human activity. The geometric nature of the spherical camera was exploited to estimate the 3D position of captured 2D poses, localising human actions to be applied to the reconstruction and affordance prediction processes. An affordance segmentation model, trained on synthetic data derived from the SUNCG dataset, maps human activity to the 3D reconstruction, providing scene affordance labels for use in future applications such as smart environments. We have demonstrated that observation of human activity in spherical video can be transferred to 2D scene affordance segmentation, to be applied to scene understanding tasks such as 3D reconstruction.

The current implementation demonstrates scene understanding from activity monitoring alone, with no direct image-based inference. Future work will investigate combining state of the art monocular 360 scene understanding approaches with the current implementation, to enhance segmentation and reconstruction results.

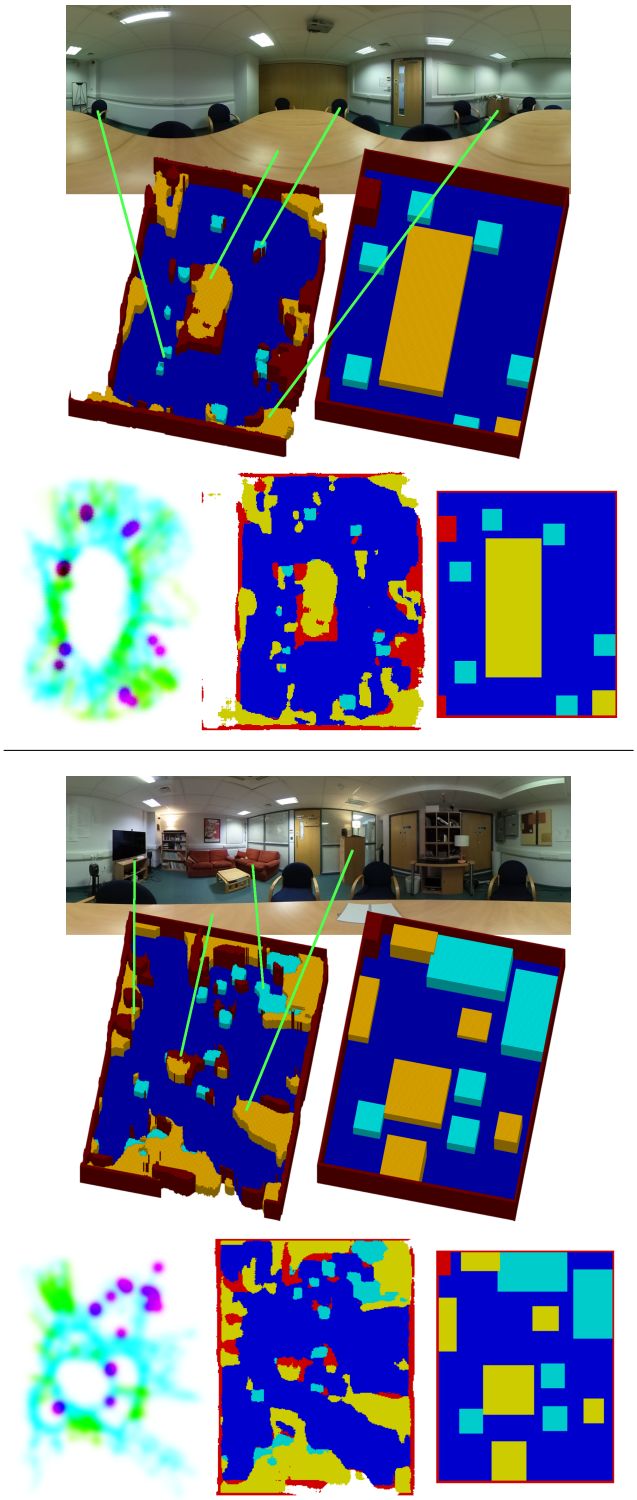


Figure 6. More 3D reconstruction and affordance segmentation results. Constructed and ground truth reconstructions (middle); activity map, predicted affordance and ground truth affordance (bottom).

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.

References

[1] Ricoh Theta S. https://theta360.com/uk/about/theta/s.html. Accessed: 2018-05-20.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *PAMI*, 2017.

[3] L. Ballan, F. Castaldo, A. Alahi, F. Palmieri, and S. Savarese. Knowledge transfer for scene-specific motion prediction. In *ECCV*, 2016.

[4] J. C. K. Chow, D. D. Lichti, J. D. Hol, G. Bellusci, and H. Luinge. Imu and multiple rgb-d camera fusion for assisting indoor stop-and-go 3d terrestrial laser scanning. *Robotics*, 2014.

[5] V. Delaitre, D. F. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. A. Efros. Scene semantics from long-term observation of people. In *ECCV*, 2012.

[6] T.-T. Do, A. Nguyen, and I. Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *ICRA*, 2018.

[7] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic. People watching: Human actions as a cue for single view geometry. *IJCV*, 2014.

[8] S. Fowler, H. Kim, and A. Hilton. Towards complete scene reconstruction from single-view depth and human motion. In *BMVC*, 2017.

[9] D. Gallup, J.-M. Frahm, and M. Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *CVPR*, 2010.

[10] H. Grabner, J. Gall, and L. Van Gool. What makes a chair a chair? In *CVPR*, 2011.

[11] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *CVPR*, 2011.

[12] P. E. Hart, N. J. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Tansactions on Systems Science and Cybernetics*, 1968.

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

[14] H. Kim, T. de Campos, and A. Hilton. Room layout estimation with object and material attributes information using a spherical camera. In *3DV*, 2016.

[15] H. Kim and A. Hilton. Block world reconstruction from spherical stereo image pairs. *CVIU*, 2015.

[16] H. S. Koppula and A. Saxena. Physically grounded spatio-temporal object affordances. In *ECCV*, 2014.

[17] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. *PAMI*, 2016.

[18] D. Lin, S. Fidler, and R. Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *ICCV*, 2013.

[19] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[20] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, 2011.

[21] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *IROS*, 2017.

[22] L. Piyathilaka and S. Kodagoda. Affordance-map: Mapping human context in 3d scenes using cost-sensitive svm and virtual human models. In *ROBIO*, 2015.

[23] D. Rother, K. A. Patwardhan, and G. Sapiro. What can casual walkers tell us about a 3d scene? In *ICCV*, 2007.

[24] A. Roy and S. Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In *ECCV*, 2016.

[25] A. G. Schwing and R. Urtasun. Efficient exact inference for 3d indoor scene understanding. In *ECCV*, 2012.

[26] N. Silberman, L. Shapira, R. Gal, and P. Kohli. A contour completion model for augmenting surface reconstructions. In *ECCV*, 2014.

[27] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, 2017.

[28] X. Wang, R. Girdhar, and A. Gupta. Binge watching: Scaling affordance learning from sitcoms. In *CVPR*, 2017.

[29] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[30] J. Xu, B. Stenger, T. Kerola, and T. Tung. Pano2cad: Room layout from a single panorama image. In *WACV*, 2017.

[31] H. Yang and H. Zhang. Modeling room structure from indoor panorama. In *SIGGRAPH*, 2014.

[32] H. Yang and H. Zhang. Efficient 3d room shape recovery from a single panorama. In *CVPR*, 2016.

[33] Y. Zhang, S. Song, P. Tan, and J. Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *ECCV*, 2014.

[34] Y. Zhu, C. Jiang, Y. Zhao, D. Terzopoulos, and S.-C. Zhu. Inferring forces and learning human utilities from videos. In *CVPR*, 2016.

[35] Y. Zhu, Y. Zhao, and S. Chun Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *CVPR*, 2015.