PROPOSAL: Form and Function Exploration in Interior Environments from Dynamic RGBD Data

Abstract

RGBD cameras is becoming more and more popular for common users to capture the environment where they live. In this paper, we present a *form and function* exploration system to reconstruct the 3D geometry and mine furniture functions in interior environments from dynamic RGBD data. We capture a sequence of RGBD images in a long period of time, in which objects may change their positions and poses frequently. To reconstruct the geometry of the cluttered indoor scene, we first cluster the objects by their motions: static objects or dynamic objects. Object motion is computed from correspondence detection from both appearance and geometry. During the motion clustering, the interrelation between different objects can be discovered from the spatial relationship at different times.

Keywords: Interior, 3d modeling, exploration, function

1 Introduction

3D indoor scenes are popular in many applications, such as games, robotics, virtual reality, etc. Modeling indoor scenes has been attracted large amount of attentions for decades in computer graphics. Recently, many techniques have been presented to generate static indoor environments, including dense modeling from RGBD data [Henry et al. 2012; Izadi et al. 2011; Xiao and Furukawa 2012; Yan et al. 2014], combing object classification and modeling [Shao et al. 2012; Nan et al. 2012; Kim et al. 2012], and synthesizing of 3D indoor scenes from large collection of examples [Fisher et al. 2012a; Xu et al. 2013].

Comparing with static scenes, dynamic scene analysis has significant value for artists in interior design, animation making, etc. The manners of how furniture objects interact with each other and how furniture objects interact with users play a very important rule in interior design. Typically, the function (behavior) of an object is interrelated with its form (geometry). In a dynamic scene, the function of an object is reflected its different forms. However, the dynamic indoor scene analysis has not been investigated much in computer graphics.

A few techniques have been proposed to analyze the spatial relationships between objects in a cluttered scene either from a boundary model [Sharf et al. 2013] or an RGBD image [Nathan Silberman and Fergus 2012]. The former technique learns features to classify the support relationship between RGBD image patches from a training dataset and then segments and classifies the support relationship of a new RGBD image. The later technique do not use any training data. It builds a support tree to describe the supporting-supported relationship between objects. Furthermore, it learns the mobility of each object/part from various poses of the repetitive instances of the same object. However, perfectly matched meshes and the large variety are required, which is very challenging for raw RGBD images.

In this paper, we present a framework to automatically explore the object behaviour in a cluttered indoor environment from a set of RGBD images without any training data . Because the raw RGBD data is noisy and incomplete due to occlusions in a cluttered environment, it is very challenging to reconstruct a perfect 3D model for the scene. Ambiguous boundaries between objects make it nontrivial to generate clear motion representation.

Robust PCA [Candès et al. 2011] is first used to segment the dynamic objects first. The a behavior map is built to represent the dynamic spatial interactions between objects in the scene.

The contributions of our system are four-fold.

1. To the best of our knowledge, our system, for the first time, performs behavior analysis in a dynamic indoor scene from raw RGBD data taken during a long time without any database.
2. We present a global optimization framework to combine object segmentation, background completion, behavior analysis in an iterative scheme.
3. Our object segmentation method using RPCA simultaneously segment objects and complete the occluded objects in a cluttered environment.
4. We present a novel graph representation and optimization technique for behavior analysis in a dynamic scene.

2 Related Work

Many techniques have been proposed to generate static 3D indoor scenes in computer graphics. Though none of them focus on dynamic scene analysis like our system, they provide valuable reference on the underlying techniques.

Reconstruction from RGBD Data For static scenes, KinectFusion [Izadi et al. 2011] enables the real-time reconstruction by holding and moving a depth camera. For large-scale indoor scenes with multiple rooms, reconstructing a dense 3D model from the noisy and incomplete scanned range data typically involves registration of point clouds in different views and a global optimization to reduce gaps in a large scene [Xiao and Furukawa 2012; Henry et al. 2012]. Their goal is mainly to generate high-quality point clouds but without semantic analysis of the objects appear in the scene. Recently, object classification is employed to assist modeling for massive indoor scenes that containing many instances of chairs, desks, etc. [Koppula et al. 2011] first introduce the learning algorithm to understand the RGBD data of an indoor scene. To further reconstruct the 3D model for a cluttered indoor scene, 3d model databases can be used as template by searching for similar 3D model and then fitting the template to the scanned data [Shao et al. 2012; Nan et al. 2012]. [Kim et al. 2012] do not manually collect 3d models to build the database. The template model is reconstructed by scanning the same object in different configuration. Each model has an additional presentation by geometric primitives. [Shao et al. 2012] trains the class model based on geometry and appearance features to segment and label the RGBD data captured under sparse views. By learned an initial model for each class of object in indoor environments from a pre-labelled database, the model are refined progressively with user-refined segmentation results. The 3D model can be generated by placing the most similar model in the database according to the RGBD data. If objects move in a scene, they can be detected and reposed by segmented and classified based on the learned model from previously reconstructed model [Liu et al. 2014]. Different with these techniques, we pay more attentions on analyzing the object behaviors from the dynamic range data.

Reconstruction from Dynamic Point clouds Many techniques have been proposed to reconstruct the object surfaces from the range data sequences. [Wand et al. 2007] uses a *statistical framework* to reconstruct the geometry from real-time range scanning. Each frame is divided into 3d pieces. A statistical model is used to iteratively merge adjacent frames by aligning pieces and optimizing their shapes. However, some geometric artifacts remain due to structured outliers and in some boundary regions. [Chang and Zwicker 2011] presents a global registration algorithm to reconstruct *articulated 3D models* from dynamic range scan sequences. The surface motion is modeled by a reduced deformable model. Joints and skinning weights are solved in the system to register point clouds in different poses. (xuejin: We may also consider the furniture objects in indoor environments as articulated models, whose shapes under different poses can be deformed through connectors like hinge, slide, and so on.)

[Bouaziz et al. 2013] propose a new formulation of the ICP algorithm using sparse inducing norms. While it achieves superior registration result on the data with outliers and missing region, only rigid alignment is handled. [Yan et al. 2014] employ a proactive capturing by asking the user to move the objects to capture both interior and exterior of a scene. The correspondence between adjacent frames is built first then segmentation. (xuejin: However, the motion information worths more than just helping registration. It can be used for analysis of object movements and functions.)

(xuejin: 1. In comparison, the range data used in our system is captured with a large time spacing and the motion of the objects in the scene varies in a wide range. 2. Can we use image/texture data for more reliable correspondence?)

Data-Driven Furniture Layout The general way producing the layout of furniture objects is to model a set of design rules and then to optimize an energy function given constraints by individuals. [Merrell et al. 2011] formulates a group of layout guidelines in a density function according to professional manuals on furniture layout. When the user specifies the room shape and an initial arrangement of the set of furniture to be placed in the room, this system generates a number of layout suggestions by a hardware-accelerated Monte Carlo sampler. Instead of manually define the layout guidelines, the hierarchical and spatial relationships of the furniture objects can be learned from a set of examples [Yu et al. 2011]. Assembling these relationships and other ergonomic factors into a cost function, multiple arrangements can be yielded quickly by simulated annealing using a Metropolis-Hastings state search step. In these methods, manual labours are required in modeling the design rules and providing an initial layout. Fisher et al. [2012b] trains a probabilistic model for indoor scenes from a small number of examples. A variety of indoor scenes can be automatically synthesized from a few of user specified examples. Indoor scenes bring more difficulties for scene analysis because there are always many cluttered objects in different scales, shapes, and functions. A focal-driven analysis and organization framework is presented for heterogeneous collections of indoor scenes [Xu et al. 2014]. They develop a co-analysis algorithm which interleaves frequent patten mining and subspace clustering. The interrelations between objects play important role during furniture arrangement in these systems. However, the 3D scene models takes many efforts to collect for training. In comparison, our system provides an efficient framework to generate 3d model examples for many further applications.

Co-analysis of shape, functions in a large database of 3D object models With the growth of 3D shape databases on the Internet, many techniques have been proposed for co-analysis in a large shape collection of the same object category. A series of geometry processing tasks such as model segmentation, shape retrieval, and shape synthesis. Point-to-point networks are used to represent the shape correspondence between shapes [Rustamov et al. 2013]. To better explore the shape space, [Huang et al. 2014] propose a framework for computing consistent functional maps within heterogeneous shape collections. Cycle-consistency of the functional map network largely reduce the noise correspondences. Based on the continuous nature of functional maps, the proposed framework outperforms point-based representation in shape interpolation, shape retrieval and classifications for both man-made and organic shapes. Large 3d model collection can also help recovering the depth map for a single image [Su et al. 2014]. With a non-rigid registration formulation, the image is popped-up to minimize the distance between corresponding points in the image and similar 3d shapes in the database. However, all these techniques focus on the geometry characteristics within one object category. In a cluttered environment, the inter-connection between different types of objects has not been investigated yet.

RGBD image understanding Though many techniques have been proposed to model a cluttered indoor environment based on databases, little work has been done for the physical interactions between objects. [?] introduce an framework to segment the RGBD image and infer the support relationship between objects in a cluttered indoor scene. A dataset is also provided for various tasks, such as recognition, segmentation and relationship inference.

3 Overview

Our system consists of two main steps, *object identification* and *behavior analysis* as Figure 1 shows.

1. *Registration/Labeling:* Given a sequence of RGBD data captured in different times and views, we first register *all the RGBD images* to segment objects and detect motions according to their low-rank characteristics using Robust PCA [Candès et al. 2011]. Given the assumption that there are typically static objects, such as wall, floor and so on, moving objects can be treated as sparse noise and the static backgrounds can be separated and completed as low-rank part by the robust PCA. The point clouds are then divided into two categories: static objects and dynamic objects. Combing the separated sparse part and the low-rank part, *object segmentation* is performed by combing the motion, depth, and appearance features in the RGBD images.

2. *Function Analysis/Behavior Map:* With the segmented regions, objects should be identified according to their appearance and geometry features among the entire image sets. We build a dynamic behavior map including all of the objects in the scene. Each node is an object/a part of an object. The edges in the graph describe the spatial relationship between objects/parts. The object behavior can be explored from its surrounding objects in the dynamic structure graph. With a dynamic behavior map, the RGBD data is re-segmented and re-labeled to obtain semantical consistency of the scene.

4 Problem

Given a set of RGBD images $\{I_i\}_{i=1,\ldots,T}$ of an indoor scene taken under different times, where each pixel in I_i is defined by $\mathbf{x} = [r,g,b,d]^T$. Our behavior analysis algorithm aims to recover all the

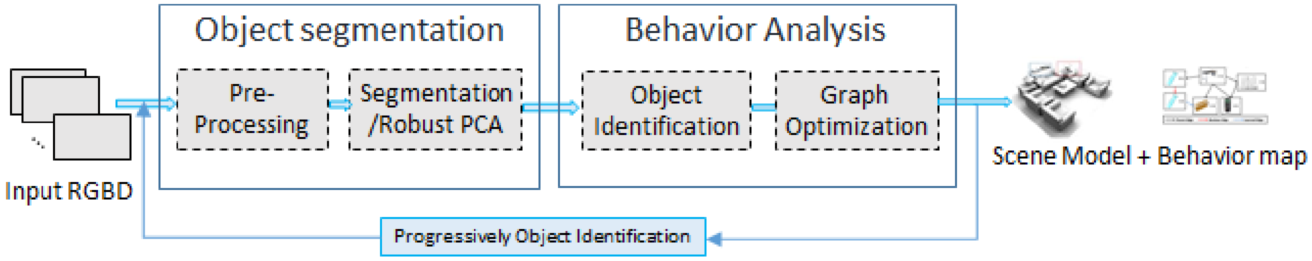

Figure 1: *System overview.*

objects $\{O_i\}_{i=1,\dots,N_o}$ appear in the scene and the behavior map M of the scene. Each object O_i is represented by its image regions $M_i, \{R_{i,t}\}_{t=1,ldots,T}$, where M_i is its registered point cloud, and $R_{i,t}$ denotes its corresponding image region in I_t. $R_{i,t} = \Phi$ if this object does not appear in the image I_t at time t.

Generally, this is a labeling problem. Each pixel at each time should be assigned with a object label. The challenging is:

1. The number of objects N_o is unknown.
2. Discrete labeling problem is NP-hard. Considering all the pixels, the node number would be $Width \times Height \times T$. The computational complexity is extremely huge.

In order to reduce the computational complexity, we need a hierarchical/progressive method to build object correspondences.

1. From the discrete RGBD pixels, we first decompose the input images into static background and moving objects using RPCA.
2. Segment each frame into multiple regions.
3. Build correspondence between regions of all the frames to identify objects in the scene.

5 Segmentation and Registration of RGBD images

Since the method used in [Shao et al. 2012] requires a certain amount of user interactions, we first register all the RGBD images taken in different views and different times using Robust PCA [Candès et al. 2011] without any pre-labeled or training data.

No work has been done for solving geometry problem using low-rank matrix. For rigid objects, small objects with large motion can be detected as noise. The challenging problem is how to detect outliers using the low-rank technique. Outlier means the objects which is not appear at all the frames. It probably appears at the begin and then disappears at the last frames. Some objects which do not show up at the begin but then appear at last frames are also outliers in low-rank problem. We have to think about how to build scene correspondences with outliers .

Segment static objects Using RPCA, we input a matrix $D_{WH\times T}$. Each column of D is a stacked image of WH pixels. After RPCA, a low-rank matrix A and a sparse matrix E is generated. Therefore, each input image can be decomposed as $I_i = I_i^b + I_i^e$.

Figure 2 shows three groups of experiments on the RPCA algorithm. Three scenes with different range of object motions are tested.

1. For the dynamic scene with large object motions , the recovered low-rank matrix A represent the static background very well while the recovered sparse matrix E represent the moving objects.
2. For the dynamic scene with moderate object motions , the recovered low-rank matrix A contains both the static background and a part of moving objects due the their overlapping at different times. The recovered sparse matrix E only indicates the part of the moving object where the depth data changes.
3. For the dynamic scene with small object motions , the result is similar with the case of moderate object motions.
4. In the three cases, new object can be correctly detected.

From the extracted A, we hope to segment the background first. Then we segment the moved objects from I_i^e or what we see about an object in the input image $I_i^b + I_i^e$ at where $I_i^e \neq 0$.

Problem of this stage:

1. Input: $D = A + E$. For each input image, it can be decomposed to $I_i = I_i^b + I_i^e$.
2. Output: $\{R_{i,t}\}_{i=1,\dots,N_o;t=1,\dots,T}$. Determine the optimal number N_o of objects in the scene. Segment the object regions in each frame. If possible, build initial object correspondence at this stage.

We segment the images A and I_i^e separately.

Why not co-segmentation Most of the state-of-the-art co-segmentation techniques segment the instances of one object in a group of images by recovering the similar appearance features. In our cases, the furniture objects usually have similar appearances and there are more than one types of objects (like chairs, cups, books) in the scenes. Moreover, the co-segmentation techniques could not complete the background/static regions due to occlusion.

6 Behavior Analysis

Object behaviour in an indoor environment is defined as the relationship between objects, including spatial relationships, function interactions and so on. We construct a graph $G = V, E$, where $V = \{V_{ij}, F_{ij}\}_{i=1,\dots,N;j=0,\dots,M}$, N is the total number of all the objects in the dynamic scene, M is the total number of RGBD frames under different times. Each node V_{ij} indicates the i^{th} object's data in the j^{th} frame. We use $F_{ij} = 0, 1$ to indicate whether the i^{th} object is visible in the j^{th} frame. Figure 3 demonstrates the final graph to represent the object correspondences and the relationship between different objects in a cluttered environment. All the red nodes are

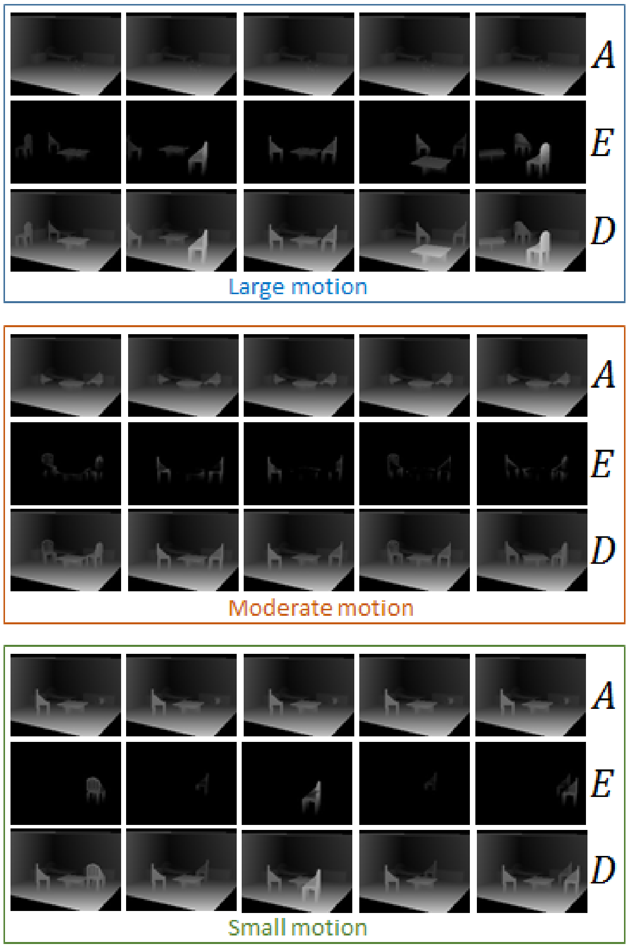

Figure 2: *Image decomposition by RPCA for three scenes.*

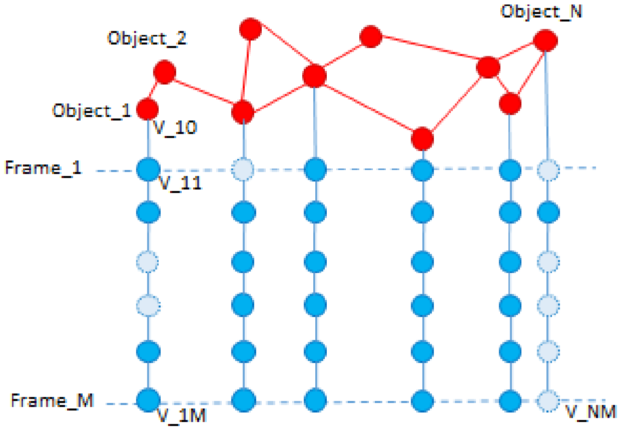

Figure 3: *Behavior graph: build the correspondence at different frames and the relationship between different objects.*

the final modeled objects in 3D space. All the blue nodes are the counterparts of each object in all the frames. If an object does not appear at frame j, $T_{ij} = 0$ and is represented by a dash blue node.

The blue edges in Figure 3 describe the correspondence of an object in different frames. The red edges describe the spatial relationships between different objects in the environment.

Initialization After we segment all the object candidates in all the RGBD images, we do not have clear blue edges to indicate accurate object correspondences. There are probably edges between each pair of nodes. Our goal is to figure out the final graph as shown in Figure 3.

The problem can be treated as a labeling problem. First, for each node V in a frame, we should assign a label $L = 1,...,N$ to indicate which object it is. This is the correspondence step. The unknowns are the number of objects before we have the correspondence. Based on graph cut theory, we can discover the proper N from the eigenvalues of an affine matrix. The affine matrix only takes the similarity between each pair of nodes into account. We should think about how to add the relationship between differen objects into the formulation.

7 Discussions

In our framework, no training data is required and the entire process is fully automatic. We do not claim that our system is able to detect all the small objects and accurately recover its motion behavior using the current automatic framework. However, using our method, we can point out which object probably appears or disappear suddenly or it probably has unusual behavior. With this kind of report, the user can easily interact with the large set of objects and rgbd images in a cluttered environment rather that manually check each frame and each object, which is non-trivial for common users.

8 Plan

1. Nov15 Test low rank algorithms. Run the code, get familier with Yi Ma's paper, define the our problem. [Done]
2. Nov19 Test Real data. Segment/Register Objects.
3. Nov25 Behavior analysis. For the graph representation, define the edge term, node term, the optimization problem.

4. Nov30 Get the initial result on behavior analysis.
5. Dec06 Siggraph Asia, discussion current results.
6. Dec15 Tests on synthetic data.
7. Dec30 Refine algorithms on real data.
8. Jan20 Paper writing and demo video.

References

Bouaziz, S., Tagliasacchi, A., and Pauly, M. 2013. Sparse iterative closest point. *Computer Graphics Forum (Symposium on Geometry Processing) 32*, 5, 1–11.

Candès, E. J., Li, X., Ma, Y., and Wright, J. 2011. Robust principal component analysis? *Journal of the ACM (JACM) 58*, 3, 11.

Chang, W., and Zwicker, M. 2011. Global registration of dynamic range scans for articulated model reconstruction. *ACM Transactions on Graphics 30*, 3.

Fisher, M., Ritchie, D., Savva, M., Funkhouser, T., and Hanrahan, P. 2012. Example-based synthesis of 3d object arrangements. *ACM Trans. Graph. 31*, 6 (Nov.), 135:1–135:11.

Fisher, M., Ritchie, D., Savva, M., Funkhouser, T., and Hanrahan, P. 2012. Example-based synthesis of 3d object arrangements. *ACM Trans. Graph. 31*, 6 (Nov.), 135:1–135:11.

Henry, P., Krainin, M., Herbst, E., Ren, X., and Fox, D. 2012. RGB-D mapping: Using kinect-style depth cameras for dense 3D modeling of indoor environments. *International Journal of Robotics Research (IJRR) 31*, 5 (April), 647–663.

Huang, Q., Wang, F., and Guibas, L. 2014. Functional map networks for analyzing and exploring large shape collections. *ACM Trans. Graph. 33*, 4 (July), 36:1–36:11.

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., and Fitzgibbon, A. 2011. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *ACM Symposium on User Interface Software and Technology*.

Kim, Y. M., Mitra, N. J., Yan, D.-M., and Guibas, L. 2012. Acquiring 3d indoor environments with variability and repetition. *ACM Transactions on Graphics 31*, 6, 138:1–138:11.

Koppula, H., Anand, A., Joachims, T., and Saxena, A. 2011. Semantic labeling of 3D point clouds for indoor scenes. In *Conference on Neural Information Processing Systems (NIPS)*.

Liu, Z., Tang, S., Xu, W., Bu, S., Han, J., and Zhou, K. 2014. Automatic 3D indoor scene updating with rgbd cameras. *Computer Graphics Forum (Pacific Graphics) 33*, 7.

Merrell, P., Schkufza, E., Li, Z., Agrawala, M., and Koltun, V. 2011. Interactive furniture layout using interior design guidelines. *ACM Trans. Graph. (Siggraph'11)*.

Nan, L., Xie, K., and Sharf, A. 2012. A search-classify approach for cluttered indoor scene understanding. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2012) 31*, 6.

Nathan Silberman, Derek Hoiem, P. K., and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *ECCV*.

Rustamov, R. M., Ovsjanikov, M., Azencot, O., Ben-Chen, M., Chazal, F., and Guibas, L. 2013. Map-based exploration of intrinsic shape differences and variability. *ACM Trans. Graph. 32*, 4 (July), 72:1–72:12.

Shao, T., Xu, W., Zhou, K., Wang, J., Li, D., and Guo, B. 2012. An interactive approach to semantic modeling of indoor scenes with an rgbd camera. *ACM Trans. Graph.*, 136–136.

Sharf, A., Huang, H., Liang, C., Zhang, J., Chen, B., and Gong, M. 2013. Mobility-trees for indoor scenes manipulation. *Computer Graphics Forums 32*, 1–13.

Su, H., Huang, Q., Mitra, N. J., Li, Y., and Guibas, L. 2014. Estimating image depth using shape collections. *ACM Trans. Graph. 33*, 4 (July), 37:1–37:11.

Wand, M., Jenke, P., Huang, Q., Bokeloh, M., Guibas, L., and Schilling, A. 2007. Reconstruction of deforming geometry from time-varying point clouds. In *Proceedings of the Fifth Eurographics Symposium on Geometry Processing*, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, SGP '07, 49–58.

Xiao, J., and Furukawa, Y. 2012. Reconstructing the world's museums. In *Proceedings of the 12th European Conference on Computer Vision*, ECCV '12.

Xu, K., Chen, K., Fu, H., Sun, W.-L., and Hu, S.-M. 2013. Sketch2scene: Sketch-based co-retrieval and co-placement of 3d models. *ACM Transactions on Graphics 32*, 4, 123:1–123:12.

Xu, K., Ma, R., Zhang, H., Zhu, C., Shamir, A., Cohen-Or, D., and Huang, H. 2014. Organizing heterogeneous scene collection through contextual focal points. *ACM Transactions on Graphics, (Proc. of SIGGRAPH 2014) 33*, 4, to appear.

Yan, F., Sharf, A., Lin, W., Huang, H., and Chen, B. 2014. Proactive 3d scanning of inaccessible parts. *ACM Transactions on Graphics(Proc. of SIGGRAPH 2014) 33*, 4.

Yu, L.-F., Yeung, S. K., Tang, C.-K., Terzopoulos, D., Chan, T. F., and Osher, S. 2011. Make it home: automatic optimization of furniture arrangement. *ACM Trans. Graph. 30*, 4, 86.