Work Surface Arrangement Optimization Driven by Human Activity

Jingjing Liu*
Beijing Institute of Technology

✉Wei Liang†
Beijing Institute of Technology

✉Bing Ning ‡
Beijing Institute of Fashion Technology

Ting Mao§
Beijing Institute of Technology

Figure 1: Our approach captures human habitual behaviors of interacting with objects on the work surface via a Hololens helmet (Left). The habitual behaviors are subsequently applied to optimize the arrangement of the work surface, resulting in a personalized arrangement suggestion for the user. The optimized arrangement is augmented in the real scene and highlighted by green box (Right).

Abstract

In this paper, we aim at guiding people to accomplish a personalized task, work surface organizing, in mixed reality environment, which can also be applied to intelligent robots. Through the cameras mounted in a MR device, e.g., Hololens, we firstly capture a person's daily activities in real scene when he uses the work surface. From such activities, we model the individual behavior habits and apply them to optimize the arrangement of the work surface. A cost function is defined for the optimization, considering general arrangement rules and human habitual behavior. The optimized arrangement is suggested to the user by augmenting the virtual arrangement on the real scene. To evaluate the effectiveness of our approach, we conducted experiments on a variety of scenes.

Index Terms: Human-centered Design—Mixed Reality—Work Surface Design—Remodeling

1 Introduction

Object arrangement on a surface is essential for a variety of scenarios, while it is hard for others to do it satisfactorily except professionals. A good object arrangement takes many factors into consideration, including object attributes, human factors, object relationships and so on. To organize the object neatly, many empirical rules are developed and to guide users, e.g. grouping objects with similar category, putting objects together with similar size [39,46]. However, some personal habitual behavior is different for different people, which is hard to be considered in general rules, such as preferences of placing objects, order of object using. For example, one person might want to place a cup on the left side of a desk instead of on the right side. Though both arrangements of the cup (on the right or the left) are reasonable, people's experiences in these two arrangements are different. Due to the diversity of habitual behavior, some may feel convenient, while others do not.

To make the work surface more tidy, people who are not good at organizing objects will invite some organizer to help them. Professional organizers sum up the user's habitual behaviors after observing or communicating with the user about his daily interactions. With these habitual behaviors, object arrangement is optimized and more user-friendly. But the effectiveness of the results depends on the experience of organizer. Moreover, it takes the organizer a lot of time on observation and communication with the user.

Recently, some automatic scene synthesis approaches [11, 34, 48, 54] have been proposed, they considering classic design rules and user personal preference. Their main layout optimized object is furniture. However, we aim to optimize the arrangement of objects on a work surface and address the problem by modeling the user's habitual behavior obtained in a real scenes. To capture and understand the user's interaction activities in the real world, we used a mixed reality system to implement our framework, as shown in Fig.1(Left). In addition, after the optimization of the arrangement, a suggested layout is presented to the user through the MR device, which helps the user to implement the arrangement in a short time.

Different people have different habitual behaviors, reflected by user's interactions with objects, e.g. number of interactions, duration of interactions, and location of interactions. The mixed reality device(Hololens) can record a users' daily interactions in real life, without disturbing the user's daily work. To model human habitual behavior and the organizing rules, we define a cost function, which measures how well the object arrangement is for the user. We apply an optimization algorithm to search for the desired arrangement, and the optimized arrangement is augmented through hololens as Fig.1(Right).

*e-mail: 3120181009@bit.edu.cn
†corresponding author, e-mail: liangwei@bit.edu.cn
‡corresponding author, e-mail: ningbing@bift.edu.cn
§e-mail: sdtzmtt@163.com

2642-5254/21/$31.00 ©2021 IEEE
DOI 10.1109/VR50410.2021.00049

In this paper, we introduce an approach for optimizing the object arrangement on a work surface driven by human habitual activities in daily life, which can recommend a more suitable arrangement for users or guide robot to organize the table intelligently. Additionally, some virtual game,e.g., Job Simulator [22] can also use the proposed idea to initialize the object arrangement for each gamer. We take office desk as an example work surface to demonstrate our approach and we also apply our approach to other work surfaces in the experiment section.

The major contributions of our work are summarized as follows:

- Propose a pipeline to optimize the object arrangement by learning habitual behavior from human activities in real life.
- Devise a cost function to model the habitual behavior and spatial constraints. The cost function is optimized to output a personalized work surface arrangement.
- Validate the effectiveness of the proposed approach by experiments and perceptual user studies on different settings.

2 Related Work

In this section, we provide a succinct overview of the scene understanding in mixed reality, and review the previous works in automatic arrangement synthesis and human activity modeling for different researches.

2.1 Scene Understanding in Mixed Reality

In the last period of time, with the development of consumer-grade mixed reality devices, there is a sharp increase in mixed reality evaluation and rendering research [20]. Mixed reality has been used for various fields, including training [6, 19], teaching [4, 8] and placement optimization [17, 26, 33]. Merenda et al. [28] explored an augmented reality interface design approached for goal-directed and stimulus-driven driving tasks. A.W.W.Yew et al. [53] proposed a teleoperation system for maintenance robots using an mixed reality environment as a human-robot interface. Liu et al. [27] used Microsoft Hololens to address the challenging problems of diagnosing, teaching, and patching interpretable knowledge of a robot. Ran Gal [12] propose a rule-based framework for generating object layouts for AR applications. Besides, to enhance the user interaction experience of mixed reality, some methods offer a virtual agent for interacting with users [15, 32].

Recent years, deep learning techniques are widely used in the object detection. Deeper CNNs have led to record-breaking improvements in the detection of more general object categories. Region-based framework is commonly used in deep detection approaches, such as DetectorNet [43], OverFeat [41] and MultiBox [7]. Lang et al. [23] obtain the semantic information of a scene by two steps: 3D scene reconstruction to obtain the 3D model with texture of the scene and detects objects via Mask R-CNN approach [14].

To extract human activities and object relationships from the video, we need to understand the movement and interaction of objects in the scene. We calculate the position and rotation of the objects through the depth information and camera parameters. By reconstructing the process in mixed reality, our approach model human habitual behavior, which is used to optimize the object arrangement.

2.2 Automatic Arrangement Synthesis

To generate furniture arrangements, some works require manual control or intervention [16, 21, 50, 51], which makes them more complicated to operate and more time consuming. Akazawa [2] use a semantic database to store furniture relationships, which must be manually specified.Germer and Schwarz [13] manually define the parent-child relationships of each object. However, this task will become prohibitive as the number of objects grows. To automatically synthesize indoor scenes, some optimization algorithms are applied. Yu et al. [54] employed the relations learned from 3D scene datasets to optimize the layout of a room through an MCMC optimizer. Genetic algorithms are also used to solve constrained object layout problems [45, 52]. And some studies [5, 44] took visual factors into account. Many works considered a set of quantifiable design criteria [29] such as aesthetic [3], ergonomic [10, 34] and functional rules [31].

In addition to the large objects such as furniture, there are some small objects on the work surface in the scene. Peter [18] propose to achieve the placement of small and decorative objects by adding new procedures into their system and each new procedure have its own set of decoration objects and rules for positioning of objects with respect to the existing furniture configuration. Based on a few user-provided examples, Fisher et al. [9] can synthesize a diverse set of plausible new scenes by learning from a larger scene database. Ritchie [35] layout the small objects by inserting them into the scene layout. Moreover, the user preferences are considered in organizing objects through collaborative filtering [1] or generative models [49] in some research.

Compared with the previous works, we focus on automatic synthesis of the small objects on the work surface through learning personal habitual behavior.

2.3 Human Activity Modeling

With the development of personalization, human activity modeling is increasingly used in various fields, including robotics, computer vision and computer graphics. A great deal of work focus on utilizing human activities for various analysis tasks. Some works obtain human action data through some sensors or devices such as HTC Vive. And we can also get data from the real world. Real data sources are mainly divided into two categories, one is the image dataset, such as Microsoft COCO (Common Objects in Context) database [25], the other is the video dataset.

There are many ways to capture and model human activities. Savva et al. [40] synthesize interaction snapshots by sampling prototypical interaction graphs learned from real-world observations of human-object interactions captured with commodity RGB-D sensors. Wang et al. [49] built action plots by observing videos of everyday activities, which are used to train a generative model based on a Recurrent Neural Network (RNN). And they use it to creating dynamic virtual environments consisting of humans interacting with objects. Rui et al. [38] can simulate how scenes are altered by human actions. Moreover, Wei et al. [24] proposed a pipeline to learn personal preferences from virtual experiences. Erich et al. [36, 37] explored the augmentation of human behaviors by building a framework of hierarchies of nonverbal channels as well as analyzing patterns of communication.

MR can easily record human interactions with real objects and render virtual optimization results. In our approach, we used the Hololens to record and modeled human activities based on the interactive behavior with real objects.

3 Overview

Optimizing work surface arrangement into a realistic and personalized configuration requires to consider of various interacting factors, such as objects relationships, interactions between the user and objects. Assume that A represents an arrangement of a work surface, consisting of N objects. For the ith object, we consider its position (x_i, y_i, z_i) and orientation θ_i. Each object in the scene is represented by a simple rectangular bounding box. The position represents the center of the bounding box. The top, bottom and front surfaces are labeled. The front surface is used to define the orientation, which represents the angle between the front surface and the front of the desk. Consequently, $A = \{A_i | A_i = (x_i, y_i, z_i, \theta_i)\}$, where $i \in \{1, 2, \cdots, N\}$.

The goal of our work is to optimize the arrangement of objects on a surface based on the learned habitual behaviors. The habitual

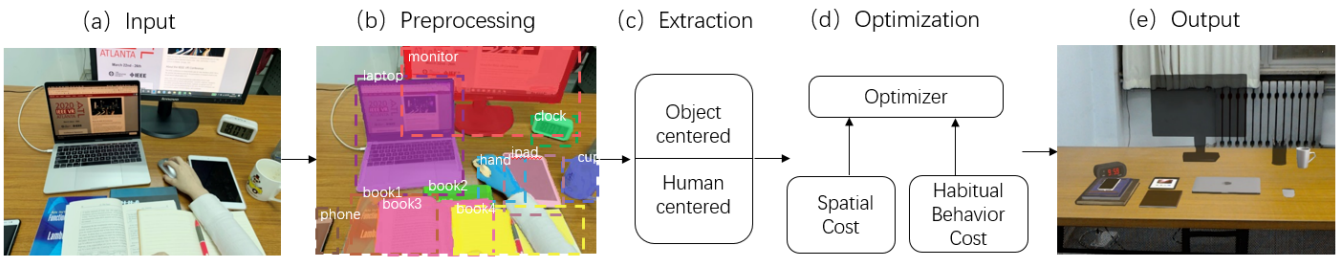


Figure 2: Overview. Capture the human activities via the HoloLens (a), from which we obtain the basic information for each object in the interaction, e.g., category, location (b). We extract object centered habitual behavior and human centered habitual behavior (c). Through the optimization which is made up of spatial cost and habitual behavior cost (d), we obtain an updated arrangement (e).

behavior refers to how a user uses objects, including the number and duration of the user's interactions with objects. To achieve this goal, the framework of our approach consist of three parts: preprocessing, habitual behavior extraction and optimization, which is shown in Fig. 2.

Preprocessing. A user do their daily work with the objects while wearing Hololens to capture his activities, which can also provide the depth information for us. To refine the user behavior, the basic information is necessary, including what and where the objects are. We first detect the objects in the scene via Mask R-CNN [14]. Then combined the detection result and depth information, the positions of the object and the user's hands are obtained. We will discuss the details in Sec. 4.

Habitual behavior extraction. The habitual behaviors are usually reflected by daily activities. From the captured daily activity, we extract the prior of object and human elbow room, including the visit frequency, the distribution of object, the human interactive sphere. The habitual behaviors are then used to optimize work surface arrangements. We will discuss the details in Sec. 5.

Optimization. A cost function is proposed to consider the spatial constrain and habitual behaviors, ensuring the rationality of the arrangement and less interactive energy caused by human activities, respectively. It will be discussed the details in Sec. 6.

4 Preprocessing

In this section, we discuss how the user interacts in real and how to obtain the projection of each object (including what and where the objects are in the 3D scene).

We take the office desk as an example to demonstrate our approach. More results of other work surfaces are shown is Sec. 7.4.

4.1 Interaction

On the office desk work surface, we define 18 object categories: book, bottle, calculator, cup, PC, glass, ipad, pen container, keyboard, keys, lamp, laptop, mobile phone, monitor, mouse, photo frame, staple, and telephone. We provide real objects of these categories to users and build up corresponding 3D virtual models. First, the user selects the objects that users daily from the real objects we provide. Then the user places the selected objects on the desk according to their own habits. Finally, the user performs some daily interaction on a desk in real while wearing HoloLens, which can record activities and provide the depth information for us. It is worth noting that the daily interactions are not instructed. That is, how to interact with these objects on the interaction duration are decided by the user himself. We do not give the user specific instructions. The reason is that we want to give the user more flexibility so as to discover his habits.

4.2 Projection

To obtain the position of the work surface where the object is placed in a 3D scene, we scan the interior scene by spatial mapping technique of the HoloLens. It usually takes about two minutes to scan and reconstruct the interior scenes. It also depends on the size of the scene. This helps us to present the optimized arrangement to the user as well.

The position of an object is regarded as equivalent to the geometric center of each object. During the interaction, the position of each object changes, which is caused by user's habitual behaviors. For all objects, the location sequence of each object is written as $P = \{p_1, p_2, \cdots\}$, where p_t is the 3D coordinate of object at time t. To understand human activities, the coordinates of each object in each frame need to be detected and estimated, including user's hands.

In order to obtain the position and category of the objects in a 3D scene, we first used the method of Mask R-CNN [14] to detect the objects in 2D images from the video stream captured by HoloLens.

After that, we map the detected objects to the 3D scene. First, we obtain the camera's parameters(including intrinsic parameters, and extrinsic parameters) and depth data of time-of-flight sensor with HoloLens research mode. Since the depth information and RGB images come from different sensors, we need to align the depth information with RGB images using the method of Lembit [47]. In this way, we obtain the corresponding depth information of the detected object. With camera's parameters, we use pinhole camera model [42] to calculate the 3D coordinates of objects in a camera coordinate system. Moreover, HoloLens rotates as the user's head rotates during capturing. Similar to the above method, we convert the coordinates relative to the current camera coordinate system to the coordinates relative to the original camera coordinate system. After this preprocessing step, we obtain the category, position and rotation of each object in a 3D scene.

5 Habitual Behavior Extraction

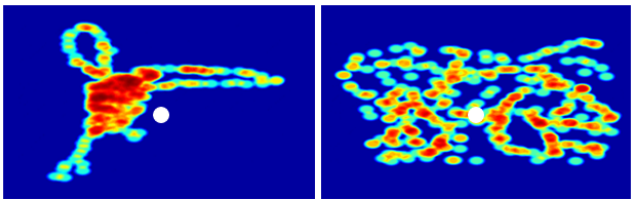
The work surface arrangement usually depends on user's habitual behaviors of using it, which can be reflected by the user's activities. Many specific relationships and distribution for objects and human interaction sphere are caused by human habitual behavior. This section describes the object prior and human elbow room based on the habitual behaviors.

5.1 Object-centered Extraction

The user's activity makes the state of each object change, from which human habitual behavior can be extracted.

Object Weight Definition Different people interact with different objects for different times, which can reflect the importance of the object to a person. However, there might be many objects in another object. For example, the user might use the pen or pencil which is in the pen container. To simplify this problem, we stipulate that it is recorded as the container instead of pen or pencil.

It is worth noting that the objects with two hands are recorded separately. Each object represents an activity. Once the interaction object is changed, the activity category changes accordingly. The



(a) (b)

Figure 3: An example of object distribution. We plot the position of (a) an ipad and (b) a cup relevant to a book (depicted by white point). The redder the color is, the more frequently it appears at that point.

total time of interaction each object is used to represent the weight of the object in the optimization.

Object Distribution Human activities cause changes in the relative positional relationship between objects, which may affect the arrangement of objects. For instance, the pairwise relationship makes two objects closer together, while mutually exclusive relationship keeps two objects away from each other. We can extract the position distribution of objects from the preprocessed video.

Take each object as the origin of coordinates and the sequence of coordinates of other objects as the target points, we generate the relative relation graph of each object with other objects, which is used to analyse the dependence between the positions of objects. As shown in Fig. 3, we take the book as an example. The positions of cup are almost uniformly distributed around the book, while the positions of ipad are more densely distributed to the left of the book.

5.2 Human-centered Extraction

The process of interaction takes place between the user and objects. Due to diversity of habitual behavior, common interaction areas and orders of different users vary. We extract habitual behavior on interaction from the preprocessed video.

Left hand-Right hand Scope. Based on the detection of user's hands in the video, we calculate the range of interactions for left hand and right hand separately, which is used to simulate the initial position of user's hands.

Object Interaction Scope. Depending on some specific properties of each object, the user has a common usage range for each one. For example, one person might prefer to interact with a book on the left of a table instead of the center of a table and the right of a table. We measure the range of interactions between individual objects, which is used to sample objects using endpoints in optimization.

Object Interaction Order. During all interaction, we respectively record the index of the objects that the user's left and right hands interact with at each time. The index sequence is written as $L = \{l_1, l_2, \cdots\}$ and $R = \{r_1, r_2, \cdots\}$, where l_t is the index of the object that the user's left hand interacts with and r_t is the index of the object that the user's right hand interacts with.

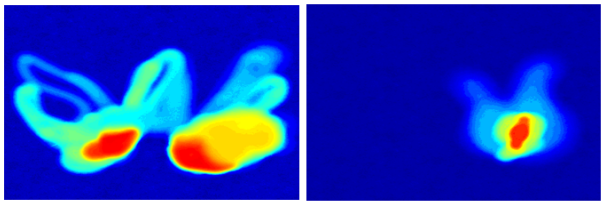
6 Optimization

We solve the problem of searching for an optimal arrangement for a user according to his habitual behaviors by an optimization process. In this section, we discuss the definition of cost function and the process of the optimization in detail.

6.1 Cost

We attempt to optimize the arrangement by analyzing the object distributions and the operating range of human interaction so that the user expends the least effort to do daily works in the updated arrangement. We define a cost function to consider the spatial constraints and habitual behavior constraints as follows:

$$C_{\text{total}}(H,A) = \omega_{\text{s}} C_{\text{s}}(A) + \omega_{\text{hb}} C_{\text{hb}}(H,A). \tag{1}$$



(a) (b)

Figure 4: An example of a user's elbow room. We use the heatmap to visualize the human's hands points (a) and the object interaction scope(b). The redder the color is, the more times the human's hands visited that point, or object interacted with at that point.

A represents an arrangement of work surface, which is explained in detail in Sec.3. H represents habitual behaviour extracted features. $C_{\text{s}}(A)$ is the spatial cost which constrains the rationality of arrangement, including the usability of each object and the compactness of the work surface. $C_{\text{hb}}(H,A)$ is the habitual behavior cost to penalize those solutions in which the user will expend more energy to complete daily work and in which the user is inconvenient to use according to his extracted habitual behavior. ω_{s} and ω_{hb} are the weights of the corresponding cost items, which are both set as 0.5.

6.1.1 Spatial Cost

In spatial cost, we consider the usability of objects and the compactness of the work surface. It is defined as follows:

$$C_{\text{s}}(A) = \omega_{\text{u}} C_{\text{u}}(A) + \omega_{\text{c}} C_{\text{c}}(A), \tag{2}$$

where $\text{C}_{\text{u}}(A)$ is the usability cost and $\text{C}_{\text{c}}(A)$ is the compactness cost. The weights ω_{u} and ω_{c} are both set as 0.5 experimentally.

Usability. In a work surface, each object should be visible and safe. Some solutions may make some objects pressed, which cause difficulties for users to find them. So we consider it is reasonable to place objects with small surface on the object with large surface, while the reverse is not. To ensure the safety of objects, we also stipulate that objects with large weight cannot be placed on light objects, and fragile objects cannot be used as parent object of other objects. We define that the parent object of an object is the object below it. For example, when a phone is on a book and the book is on the desk, the parent of the phone is the book and the parent of the book is the desk. The cost is defined as:

$$C_{\text{u}}(A) = \frac{1}{N}\sum_i \max\left(0, \frac{m_i}{m_{pr_i}} - 1, \frac{a_i \cdot b_i}{a_{pr_i} \cdot b_{pr_i}} - 1\right), \tag{3}$$

where pr_i is the parent of object i. m_i and m_{pr_i} are the mass of object i and pr_i, which are defined according to its weight in real life. a_i and b_i are the length and width of object i respectively. Similarly, a_{pr_i} and b_{pr_i} are the length and width of object pr_i.

Compactness. Another aspect that is considered in work surface arrangement is the compactness of desktop objects. Except for necessary spaces for objects on the work surface, the more compact the objects on the work surface, the cleaner it looks and the more free space the user can use. Therefore, we use the minimum coverage of all objects on the work surface to reflect the area used and calculate it from an overhead view.

We illustrate the calculation process in Fig. 5. Technically, we first detect the most marginal objects on the work surface, including left, right, top and bottom, which are marked in yellow. To simplify the calculation of the coverage area, we define the coverage area as a minimum rectangle containing all objects. This rectangle is marked in gray. We discretize the surface by grid, and count the number of gray region cells. The cost is formulated as:

$$C_{\text{c}}(A) = \frac{1}{G}\sum \mathbb{E}_{grid}, \tag{4}$$

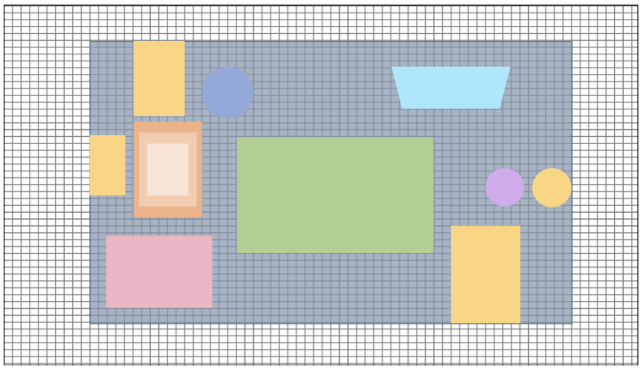


Figure 5: Compactness cost. This is the top view of the desk, and we label the most marginal object in the arrangement as yellow objects. The gray region is the minimum rectangle containing all objects and is used to evaluate the compactness of the desk.

where $\mathbb{E}_{grid}$ is an indicator function, representing the coverage area cells. G is the normalization parameter, which is the number of all cells, regardless of whether it is occupied space or not.

6.1.2 Habitual Behavior Cost

In the habitual behavior cost, we consider the dependency relationships of objects and the interactive energy cost. It is defined as follows:

$$C_{\text{hb}}(H,A) = \omega_{\text{dr}} C_{\text{dr}}(A) + \omega_{\text{in}} C_{\text{in}}(H,A) + \omega_{\text{f}} C_{\text{f}}(H,A), \quad (5)$$

where $C_{\text{dr}}(A)$ is the dependency relationships cost, $C_{\text{in}}(H,A)$ is the interaction cost and $C_{\text{f}}(H,A)$ is the free space cost. Experimentally, ω_{dr} is set 0.3, ω_{in} is set 0.3 and ω_{f} is 0.4.

Dependency Relationships. There will be some interdependencies between object and the other objects. We create a distribution of every two object. Fig. 3 shows an example. The dot represents the other object and the star is the target object. The denser the point of the other object, the stronger the correlation between the target object and the reference object in this area. To favor the dependencies, the cost increases when the displacement vectors between two objects of a strong dependence increases. Thus, the cost penalty ensures that the related object will be in a relatively fixed position, e.g., keyboard and monitor. The cost is defined as:

$$C_{\text{dr}}(A) = \frac{1}{2N} \sum_i \sum_m \mathbb{R}_{i,m}, \quad (6)$$

where i and m are object i and m. $\mathbb{R}_{i,m}$ represents the relative position of the object i and m, which can be calculated from the distribution of object i and m. As shown in Fig 3, in the distribution of i object, the denser the position of object m relative to object i in the distribution, the smaller the $\mathbb{R}_{i,m}$ is.

Interaction. In the interaction cost, we consider the movement distance and the loads carried by the user when he moves around. It is defined as follows:

$$C_{\text{in}}(H,A) = \frac{1}{N} \sum_i \frac{D(i,i')}{D_{\max}} \frac{m_i}{M_{\max}} \frac{t_i}{T_{\max}}, \quad (7)$$

where $D(i,i')$ is the distance from the position of object i in the arrangement to the interaction position. We generate the interaction position of object i by the real object interactive distribution as Fig. 4(b). $D(i,i')$ is calculated as the Euclidean distance of the two positions i and i'. m_i is the mass of the object the user carries, which is marked ahead of the optimization process. t_i is the total number of occurrences of object i in L,R. L,R is the index of the objects that the user's left and right hands interact with at each time. $D_{\max}$ is the diagonal length of the surface. $M_{\max}$ is the max mass of all objects. $T_{\max}$ is the total elements in L,R. N is the total number of objects on the work surface.

Manipulation. This cost ensures the user's use of the common interactive area. We take the user's hand interaction scope and the importance of the object into account. And we use the interactive time to evaluate the importance of objects. The cost is defined as:

$$C_{\text{f}}(H,A) = \frac{1}{N} \sum_i \frac{Q(i,h)}{t_i}, \quad (8)$$

where h is the hand, i is object i. $Q(i,h)$ represents the relative distribution between the position of hand and the position of object i, which is shown in Fig 4(a). The redder the color is, the greater the $Q(i,h)$ is. t_i is the total number Ti is the total number of interactions of object i. N is the total number of objects on the work surface.

6.2 Simulated Annealing

We search for a proper work surface arrangement by an optimization process. The goal of the optimization is to find position (x_i, y_i, z_i) and orientation θ_i of all the object that minimize the total cost function $C_{\text{total}}(H,A)$. To search for a good arrangement configuration efficiently, Simulated Annealing(SA) algorithm with the simple Metropolis criterion is applied to explore the solution space and it has the ability of jumping out from a local minimum by the strategy of accepting a worse solution by a probability.

6.2.1 Move

The SA algorithm works by choosing a "move" to generate a new solution. We define two moves to explore the solution space: position move and orientation move, corresponding to adjusting the position (x_k, y_k, z_k), the orientation θ_i and the parent object of each object. The sampling process performs these two moves alternately.

Position. During the optimization, we find the position of each object in two steps. First, we find the position in (x,y)-space. Then in the z-dimension, the object's z-position is given by generating its parent object.

In the (x,y)-space position move, the sampler generates a new position (x'_k, y'_k) based on the current location (x_k, y_k):

$$(x_k, y_k) \rightarrow (x'_k, y'_k), \quad (9)$$

Position move is the basic operation of the optimization that modifies the position of objects in a workspace. An object k is selected and its position is updated by a move. Mathematically, we define a linear operation to accomplish the translation move denoted as $(x_i + \Delta x, y_i + \Delta y)$. We sample the variation $(\Delta x, \Delta y)$ using a Gaussian distribution.

In z-dimension, the sampler generates a new object j from the object set $\{0, 1, 2 \cdots\}$ expect the i as the parent of object i. The (z'_k) is the sum of the c_j and half of the c_i. c_j and c_i are the height of object i and object j.

Orientation. In the orientation move, the sampler generates a new orientation θ_i based on the current orientation θ'_i. An object i is selected, and its orientation is updated with a rotation change $(\theta_i + \Delta\theta)$. The rotation change $\Delta\theta$ is generated from a Gaussian distribution.

It is worth noting that we constrain the solution space into the actual work surface. Moreover, in the sampling process, we performs the collision detection for each object in the scene. If a collision occurs, another random position, orientation or parent object will be sampled.

Figure 6: Work surface arrangement optimization visualization. As the optimization process proceeds, the object configuration is iteratively changed until it achieves an optimized final arrangement.

6.2.2 Sampling

Given an initial configuration of a work surface A_0, our approach explores the solution space in the following two steps. We firstly propose a move $A^i \rightarrow A^*$. The new move is accepted or rejected according to the Metropolis-Hastings, which can be formulated as:

$$\alpha(A^i, A^*) = \min\left\{\frac{p(A^*)}{p(A^i)}, 1\right\}, \tag{10}$$

$p(\cdot)$ is computed by the defined cost:

$$p(\cdot) = \frac{1}{Z}\exp\frac{-C_{\text{total}(\cdot)}}{Temp}, \tag{11}$$

The temperature of the annealing process is donated by $Temp$ and Z is a constant value for normalization. To explore the solution space in a more aggressive way, the $Temp$ is initialized with a large value at the early iteration, which keeps decreasing in the optimization. And then, the $Temp$ is fixed with a small value close to 0, which makes it possible that the solution can be well-refined by the sampler. Specifically, the $Temp$ is initialized at 1.0 and decrease it by 0.05 for every 10 iteration. The optimization will be terminated if the total cost value is less than 5% over the past 30 iterations. The full optimization takes about 350-500 iterations in our experiments. Fig.6 shows an example optimization process.

7 Experiments

7.1 Implementation

We implemented our approach using C# and Unity 2018. Due to the limited computing power of the Hololens, we ran the optimization approach on a PC equipped with 32GB of RAM, a Nvidia Titan X graphics card with 12GB of memory, and a 2.60GHz Intel i7- 5820K processor. The captured human activity and the optimized results are transmitted between the Hololens and the PC by WiFi. Based on the transmitted results, the user can properly organize the real objects of the work surface according to the virtual optimization result via the HoloLens mixed reality headset.

7.2 Qualitative Experiment

To evaluate the effectiveness of our optimizing approach qualitatively, we compare with other approaches:

Our Approach: Our approach generates an optimization results after learning the users' habitual behavior, which is presented by the virtual objects corresponding to the real object of the model library. The desks were organized according to the results.

Consultant Approach: The desks were manually organized by an organising consultant. We recruited a consultant who had about 3 years experiences of helping people organizing objects and didn't know any participants before. He communicated with each participant to learn their habits and then put the objects in order.

User Approach: The desks were manually organized by the participants before they experienced in the scene.

Participants. 24 participants were recruited who were unaware of the purpose of the user study. The participants included 15 males and 9 females whose ages ranged from 18 to 50. All the subjects reported normal or corrected-to-normal vision with no color-blindness. 13 subjects reported that they have little experience in organizing things. 4 subjects reported that they organized their things every three or four months, and the other 7 subjects organized every day. 16 subjects reported that they did not have any experience in VR/AR devices using. The participants were randomly divided into 3 groups. For each group, 8 participants were randomly assigned to each of the 3 traditional desk types: rectangular, square and right angle.

Metrics. We define three metrics to investigate the comparison:

Convenience: to evaluate how comfortable the user feels when he moves his hands among the objects;

Harmony: to evaluate the cleanliness level the user feels about the arrangement of objects;

Overall: to evaluate the overall feeling about desk-top arrangement.

The score range of each metric is from 0 to 10. We instructed all participants about the metrics and encouraged them to ask questions. During the evaluation, the order for the desks generated by different approaches is random.

Procedure. The user study was carried out via a HoloLens. Before the evaluation, each participant went through a 5-minute warm-up section so that they got familiar with using the HoloLens helmet. A warm-up program is used to help the user to learn about the basic operation via the HoloLens, which is a MR application named Holotour [30].

After the warm-up process, the participant chose objects from the whole object set discussed in Sec.4.1 and placed the selected objects on the desk by themselves. Then he wore HoloLens to do their daily work on the desk-top. We captured the video sequence of users' interaction with objects for a period of time, which is at least three hours per user. We optimized the arrangement of objects on the desk-top based of the captured user activities.

We performed the user study on three scenes, which were created concerning the traditional shape of the desk (square, rectangle and right angle).

Results and Analysis. We compare the results generated from the three approaches by the mean and standard deviation of the users' ratings. We also carried out ANOVA tests of the scores to analyze the significance. Fig 7 shows the average scores and their standard deviation of the harmony, convenience and overall.

For the harmony rating item, our approach got the highest score ($M = 7.33, SD = 0.89$), followed by the approach of Consultant ($M = 7.04, SD = 1.06$) and User ($M = 6.25, SD = 0.97$). The ratings of our approach and the Consultant on harmony are close. Based on the organizing consultant has the basic knowledge of design, it is better to arrange the degree of harmony than the user himself. It means that our approach takes into account the principles of design harmonization such as compactness.

For the convenience rating term, User($M = 7.67, SD = 0.75$) is the best compared to the approach of Consultant($M = 6.08, SD =$

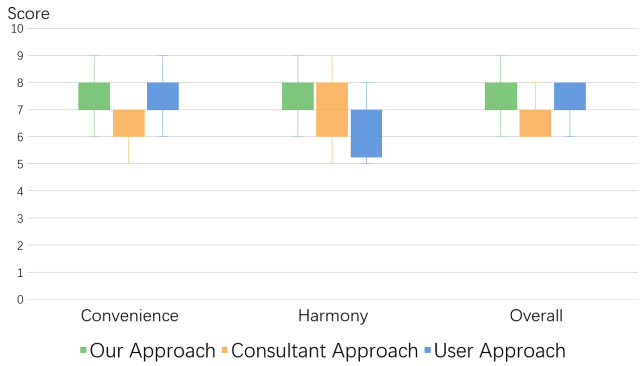


Figure 7: The box plot of the user rating on the work surfaces generated by ours, by the consultant and by the user approach.

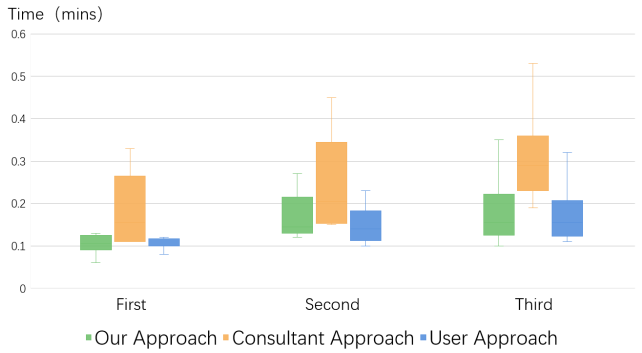


Figure 8: The box plot of quantitative set-up time for setting up the work surfaces generated by ours, by the consultant and by the user approach before user do the first, the second or the third frequent task in daily life.

0.81) and our approach($M = 7.63$, $SD = 0.81$). We interviewed the consultant's personalized approach to designing object arrangement He said he arranged the objects combined with the personal behavior preferences by interviewing the user. It may ignore some details of the user interaction with objects, such as the custom interaction point between the user and the object. However, the user may know in his own subconscious.

For the overall rating term, our approach ($M = 7.71, SD = 0.73$) also performs better than the approach of Consultant ($M = 6.88$, $SD = 0.73$) and User ($M = 7.13$, $SD = 0.67$). It indicates that recorded habitual behaviors may help the user to organize his work surface and generate a personalized work surface arrangement.

There is a significant difference among three approaches for harmony($\chi^2 = 8.27$, $p < .05$, $df = 2$), convenience($\chi^2 = 28.79$, $p < .05$, $df = 2$) and overall($\chi^2 = 10.17$, $p < .05$, $df = 2$). We also did an ANOVA test on the harmony, convenience and the overall rating at the $\alpha = 0.05$ significance level. In the harmony, the results showed statistically significant between our approach and the User approach($F_{[1,23]} = 3.94$, $p < .05$). In the convenience, the results showed statistically significant between our approach and the consultant approach($F_{[1,23]} = 0.11$, $p < .05$). In the overall, our approach showed statistically significant with the Consultant approach ($F_{[1,23]} = 2.13$, $p < .05$) and the User approach ($F_{[1,23]} = 2.07$, $p < .05$).

The results did not show statistically significant between our approach and the Consultant approach in harmony($F_{[1,23]} = 1.42$, $p = 1 > .05$), which shows that the spatial layout costs in our approach work. And our approach did not have statistically significant with User approach in convenience($F_{[1,23]} = 3.95$, $p = 0.58 > .05$). It shows that our approach can understand the human behavior and learn their habitual behavior.

It is interesting that the optimization results given by our approach for a rectangular table are better for users than for rectangular and square tables. In a larger desk, there are more types of objects selected by the user and the more difficult it is for users to clean up. The more complex the user interaction, the harder the user habits captured by the conversation.

7.3 Quantitative Experiment

We verify the effectiveness of our optimizing approach quantitatively by the time of each approach takes discussed in Sec.7.2 and how long it takes each participant to set up the desktop to do the three most frequently done tasks in the different optimization results.

Participants. 24 participants were the same participants as in Sec.7.2. They were randomly divide into 3 groups again. For each group, there are 8 participants and they prepare for one of the three most frequently completed tasks.

Procedure. We recorded the processing time for each approach during the experiment. For Our approach, it is the time it takes the algorithm to do it. For the Consultant, it refers to the time the consultants communicate with the user. For the User, it is the time of the user's initialization.

The participants were asked to fill out three tasks they usually do in the order of their frequency such as writing, reading and so on. Then the participants needed to set up the desktop for one of the three tasks based on the optimization result in three methods of the qualitative experiment. At the same time, we recorded the set-up time. To ensure the accuracy of the calculation, we calculated the time by recording the process on video.

Results and Analysis. For the processing time for each approach, the Consultant took the longest processing time. Our approach was accomplished within an average time of 3.2 minutes (the shortest is about 1.7 minutes, and the longest is about 4 minutes), while the Consultant approach spent 38.42 minutes on average (the shortest is about 28 minutes, and the longest is about 65 minutes). The User approach spent 4.34 minutes on average (the shortest is about 1.9 minutes, and the longest is about 6.8 minutes).

To demonstrate that our approach can learn the habitual behavior, we recorded the time for setting up the desktop for the tasks that users often do. Fig.8 shows the time of each participant set up the desk to do frequently completed tasks. In each level of the task, the desktop are set up faster using in Our approach's result (first: $M = 0.11$ minutes, $SD = 0.04$ minutes; second: $M = 0.17$ minutes, $SD = 0.05$ minutes; third: $M = 0.18$ minutes, $SD = 0.07$ minutes) than the Consultant's results(first: $M = 0.18$ minutes, $SD = 0.08$ minutes; second: $M = 0.24$ minutes, $SD = 0.11$ minutes; third: $M = 0.31$ minutes, $SD = 0.1$ minutes), which have no significant differences with the User's results(first: $M = 0.12$ minutes, $SD = 0.04$ minutes; second: $M = 0.15$ minutes, $SD = 0.04$ minutes; third: $M = 0.18$ minutes, $SD = 0.06$ minutes). It is interesting to see that as the user's level of frequent doing declines, the less obvious the user's habits become, and the more difficult it is for consultants to better capture the user's habits.

The quantitative results indicate that our approach can generate the suitable work surface arrangement in a short time, which can learn the user's habitual behavior intelligently.

7.4 Extended Experiment

To investigate the generality of our approach, we experiment on two other work surfaces: kitchen plane and dresser surface. There are different objects need to be optimized in different scenes. For the kitchen plane, the objects we optimized are various flavor jars, cookers and so on. For the dresser surface, the objects are skin care products, make-up tools and so on.

Fig. 9 shows the optimized results. For the kitchen plane, the tissue box and the flavor jar were optimized close together, which was more consistent with the user's habits of interaction. More specifically, he frequently used to wipe paper towels around the flavor jars bottle. In addition, the user preferred to use his left hand to take out the cutting board. In the optimized arrangement, the cutting board was optimized to the left side of the surface. For the dresser surface, the user preferred to use a powder puff to apply skincare products of her face. When the user started to make-up, she took a powder puff from the toolbox, and she moved the skincare products to the front of the table. The user used powder puff to dip into skin care products. Comparing the initial arrangement with the optimized arrangement, we can see that the optimized arrangement brought the skincare products and the powder puff closer. Besides, the vase was used for decoration and the user preferred to use her right hand, so it was moved from the right side to the left side in the optimized layout.

Our approach can be applied to more than one workspace category. The results of the experiment show that our approach can learn the habitual behavior from the user's interaction in various work surface and optimize a reasonable arrangement of the objects.

8 Summary

In this paper, we present an approach to remodel work surface arrangement based on the habitual behaviors. A new pipeline is designed for better learning the habitual behaviors from the interaction between users and objects. Based on the learned habitual behaviors, we define a cost function to evaluate how reasonable the layout of the arrangement is and how comfortable the user may feel about. In the cost function, we consider the habitual behaviors and spatial constrains. The cost function is optimized to output an updated personalized work surface arrangement by a SA algorithm, which is more suitable for the user.

Additionally, our approach benefits a lot from the mixed reality. The users are allowed to interact with objects in the real-world, which is much more natural and realistic than objects in the virtual world. By capturing users' activities with Hololens, we further extract the habitual behavior for improving the arrangement and better fit the individual user, especially there is a particular preference existed. Extensive experiments are conducted for verifying the effectiveness of the proposed method. It is worth mentioning that the proposed cost function and the optimization method can be further extended to various work surfaces.

Limitations. The 3D scene reconstruction and object detection techniques provide the geometry and the semantics information for our approach to instantiate the human behaviors. The performance of our method is likely to be affected by quality of the preprocessing. For example, we can only detect a limited number of objects through Mask R-CNN. If the object was not captured or detected accurately, we marked manually. However, we believe that better detection performance will help to improve it. In addition, we ran the optimization process on a PC rather than on the Hololens due to the limited computing power of the Hololens. With further advancement in the computing capabilities of mixed reality devices, we may be able to run our whole approach on a mixed reality device efficiently.

Due to the weight of the Hololens helmet, it is difficult for the user to interact with the objects for a long time. It may reduce the collection of data on learning habitual behaviors. If Google glasses or other light devices can be used, this will be improved.

Future Work. In our work, we consider the fixed and limited objects of a work surface. One possible extension is to explore how to optimized the objects of a larger space. For example, we can apply our pipeline to optimize all objects in a room or in a house.

Another possible extension is to transfer the learned habitual behaviors into other arrangements. For example, we optimized the objects arrangements on the dresser based on the habitual behaviors learned from the user interactions captured on the desk.

Figure 9: The initial and the optimized results of a kitchen surface and a dresser surface. The work surface arrangement is optimized based on the user's daily activities.

Another future direction is to extend our approach for driving intelligent robot. Equipped with high-quality sensors, an intelligent robot can capture the user's interaction activities. Futhermore, our approach could be applied for the intelligent robot and help the robots to arranging the objects on the work surface intelligently.

In our current approach, we only consider the habitual behaviors of one person. It would be interesting to explore our approach to multiple people scenarios, where the preferences from different roles could be considered. For example, if we want to optimize an office or coffee bar, different preferences from different people should be considered to optimize the arrangement.

9 Acknowledgments

This project was supported by the National Natural Science Foundation of China(NSFC) under Grant No. 61972038.

References

[1] N. Abdo, C. Stachniss, L. Spinello, and W. Burgard. Organizing objects by predicting user preferences through collaborative filtering. *The International Journal of Robotics Research*, 35(13):1587–1608, 2016.

[2] Y. Akazawa, Y. Okada, and K. Niijima. Automatic 3d scene generation based on contact constraints. In *Computer Graphics and Artificial Intelligence*, pp. 593–598, 2005.

[3] D. K. Ballast. *Interior design reference manual*. Professional Publications Incorporated, 2002.

[4] W. L. Brandao and M. S. Pinho. Using augmented reality to improve dismounted operators' situation awareness. In *Virtual Reality*, 2017.

[5] N. R. Dayama, K. Todi, T. Saarelainen, and A. Oulasvirta. Grids: Interactive layout design with integer programming. 2020.

[6] N. Döllinger, C. Wienrich, E. Wolf, M. Botsch, and M. E. Latoschik. Vitras-virtual reality therapy by stimulation of modulated body image-project outline. *Mensch und Computer 2019-Workshopband*, 2019.

[7] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2147–2154, 2014.

[8] O. Fajarianto, M. I. Setiawan, A. Mursidi, D. Sundiman, and D. A. P. Sari. The development of learning materials for introduction of animals in early childhood using augmented reality. In *International Conference on Knowledge Management in Organizations*, pp. 722–727. Springer, 2018.

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

[10] M. Fisher, M. Savva, Y. Li, P. Hanrahan, and M. Nießner. Activity-centric scene synthesis for functional 3d scene modeling. *ACM TOG*, 34(6):179, 2015.

[11] Q. Fu, X. Chen, X. Wang, S. Wen, B. Zhou, and H. Fu. Adaptive synthesis of indoor scenes via activity-associated object relation graphs. *ACM TOG*, 36(6):201, 2017.

[12] R. Gal, L. Shapira, E. Ofek, and P. Kohli. Flare: Fast layout for augmented reality applications. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 207–212, 2014. doi: 10 .1109/ISMAR.2014.6948429

[13] T. Germer and M. Schwarz. Procedural arrangement of furniture for real-time walkthroughs. In *Computer Graphics Forum*, vol. 28, pp. 2068–2078. Wiley Online Library, 2009.

[14] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017.

[15] T. Holz, A. G. Campbell, G. M. P. O'Hare, J. W. Stafford, A. Martin, and M. Dragone. Mira-mixed reality agents. *International Journal of Human - Computer Studies*, 69(4):251–268, 2011.

[16] C. Jiang, S. Qi, Y. Zhu, S. Huang, J. Lin, L. F. Yu, D. Terzopoulos, and S. C. Zhu. Configurable 3d scene synthesis and 2d image rendering with per-pixel ground truth using stochastic grammars. *IJCV*, 126(9):920–941, 2018.

[17] Jun-Chao, Yan, Jun, Ren, Lin-Lin, Ren, Yi, Yang, Si-Fan, and Y. and. Optispace: Automated placement of interactive 3d projection mapping content. *Sensors and Actuators A: Physical*, 2019.

[18] P. Kán and H. Kaufmann. Automatic furniture arrangement using greedy cost minimization. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 491–498. IEEE, 2018.

[19] H. Kim, J. L. Gabbard, A. M. Anon, and T. Misu. Driver behavior and performance with augmented reality pedestrian collision warning: An outdoor user study. *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1–1, 2018.

[20] K. Kim, M. Billinghurst, G. Bruder, H. B.-L. Duh, and G. F. Welch. Revisiting trends in augmented reality research: A review of the 2nd decade of ismar (2008–2017). *IEEE transactions on visualization and computer graphics*, 24(11):2947–2962, 2018.

[21] K. A. H. Kjølaas. Automatic furniture population of large architectural models. *Doctoral dissertation*, Massachusetts Institute of Technology, 2000.

[22] O. Labs. Job simulator. `https://store.steampowered.com/app/448280/Job_Simulator/`.

[23] V. Lang, W. Liang, and L.-F. Yu. Virtual agent positioning driven by scene semantics in mixed reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 767–775. IEEE, 2019.

[24] W. Liang, J. Liu, y. Lang, B. Ning, and L.-F. Yu. Functional workspace optimization via learning personal preferences from virtual experiences. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):1836 – 1845, 2019.

[25] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. 2014.

[26] D. Lindlbauer, A. M. Feit, and O. Hilliges. Context-aware online adaptation of mixed reality interfaces. In *the 32nd Annual ACM Symposium*, 2019.

[27] H. Liu, Y. Zhang, W. Si, X. Xie, Y. Zhu, and S.-C. Zhu. Interactive robot knowledge patching using augmented reality. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1947–1954, 2018.

[28] C. Merenda, H. Kim, K. Tanous, J. L. Gabbard, B. Feichtl, T. Misu, and C. Suga. Augmented reality interface design approaches for goal-directed and stimulus-driven driving tasks. *IEEE transactions on visualization and computer graphics*, 24(11):2875–2885, 2018.

[29] P. Merrell, E. Schkufza, Z. Li, M. Agrawala, and V. Koltun. Interactive furniture layout using interior design guidelines. In *ACM TOG*, vol. 30, p. 87, 2011.

[30] Microsoft. Holotour. `https://www.microsoft.com/en-us/p/holotour/9nblggh5pj87`.

[31] M. Mitton and C. Nystuen. *Residential interior design: A guide to planning spaces*. John Wiley and Sons, 2016.

[32] K. Miyawaki and M. Sano. A virtual agent for a cooking navigation system using augmented reality. 2008.

[33] E. Ofek, B. Ens, N. Bruce, and P. Irani. Spatial constancy of surface-embedded layouts across multiple environments. In *SUI 2015*, 2015.

[34] S. Qi, Y. Zhu, S. Huang, C. Jiang, and S. C. Zhu. Human-centric indoor scene synthesis using stochastic grammar. In *CVPR*, 2018.

[35] D. Ritchie, K. Wang, and Y.-a. Lin. Fast and flexible indoor scene synthesis via deep convolutional generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6182–6190, 2019.

[36] D. Roth, G. Bente, P. Kullmann, D. Mal, C. F. Purps, K. Vogeley, and M. E. Latoschik. Technologies for social augmentations in user-embodied virtual reality. In *25th ACM Symposium on Virtual Reality Software and Technology*, VRST '19, 2019, to appear.

[37] D. Roth, C. Kleinbeck, T. Feigl, C. Mutschler, and M. E. Latoschik. Beyond replication: Augmenting social behaviors in multi-user virtual realities. In *Proceedings of the 25th IEEE Virtual Reality (VR) conference*, pp. 215–222, 2018.

[38] M. Rui, H. Li, C. Zou, Z. Liao, T. Xin, and Z. Hao. Action-driven 3d indoor scene evolution. *Acm Transactions on Graphics*, 35(6):173, 2016.

[39] D. Rus and P. D. Santis. *The Self-Organizing Desk*. 1997.

[40] M. Savva, A. X. Chang, P. Hanrahan, M. Fisher, and M. Nießner. Pigraphs: learning interaction snapshots from observations. *Acm Transactions on Graphics*, 35(4):139, 2016.

[41] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[42] P. Sturm. *Pinhole camera model*. 2014.

[43] C. Szegedy, A. Toshev, and D. Erhan. Deep neural networks for object detection. In *Advances in neural information processing systems*, pp. 2553–2561, 2013.

[44] K. Todi, J. Jokinen, K. Luyten, and A. Oulasvirta. Familiarisation: Restructuring layouts with visual learning models. In *the 2018 Conference*, 2018.

[45] T. Tutenel, R. Bidarra, R. M. Smelik, and K. J. De Kraker. Rule-based layout solving and its application to procedural interior generation. In *CASA Workshop on 3D Advanced Media In Gaming And Simulation*, 2009.

[46] P. J. Vaccaro. Organizing your desk. *Family Practice Management*, 2003.

[47] L. Valgma. 3d reconstruction using kinect v2 camera. *University of Tartu*, 2016.

[48] H. Wang, W. Liang, and L.-F. Yu. Scene mover: Automatic move planning for scene arrangement by deep reinforcement learning. *ACM Transactions on Graphics*, 39(6), 2020.

[49] H. Wang, S. Pirk, E. Yumer, V. G. Kim, O. Sener, S. Sridhar, and L. J. Guibas. Learning a generative model for multi-step human-object interactions from videos. In *Computer Graphics Forum*, vol. 38, pp. 367–378. Wiley Online Library, 2019.

[50] K. Wang, M. Savva, A. X. Chang, and D. Ritchie. Deep convolutional priors for indoor scene synthesis. *ACM TOG*, 37(4):70:1–70:14, 2018. doi: 10.1145/3197517.3201362

[51] K. Xu, J. Stewart, and E. Fiume. Constraint-based automatic placement for scene composition. *Proceedings - Graphics Interface*, pp. 25–34, 2002.

[52] Y.-T. Yeh, L. Yang, M. Watson, N. D. Goodman, and P. Hanrahan. Synthesizing open worlds with constraints using locally annealed reversible jump mcmc. *ACM TOG*, 31(4):1–11, 2012.

[53] A. Yew, S. Ong, and A. Nee. Immersive augmented reality environment for the teleoperation of maintenance robots. *Procedia CIRP*, 61:305–310, 2017.

[54] L.-F. Yu, S. K. Yeung, C.-K. Tang, D. Terzopoulos, T. F. Chan, and S. Osher. Make it home: automatic optimization of furniture arrangement. In *ACM TOG*, 2011.