Volume 0 (1981), Number 0 pp. 1–13 *Pre-print*

Context-Aware Mixed Reality: A Learning-based Framework for Semantic-level Interaction

L. Chen[1] , W. Tang[1†], N. W. John[2], T. R. Wan[3], J. J. Zhang[4]

[1]Creative Technology, Bournemouth University, UK
[2]Department of Computer Science, University of Chester, UK
[3]School of Informatics, University of Bradford, UK
[4]National Centre for Computer Animation, Bournemouth University, UK

(a) (b) (c)

Figure 1: A MR shooting game (a) and a throwing plates game (b) developed by the proposed framework that can provide material-specific physical interactions detected by deep semantic material learning method. (c) the semantic interaction layer in which different materials are textured by different colors (Green: glass, Purple: painted, Blue: fabric, Yellow: wood, Red: carpet).

Abstract

Mixed Reality (MR) is a powerful interactive technology for new types of user experience. We present a semantic-based interactive MR framework that is beyond current geometry-based approaches, offering a step change in generating high-level context-aware interactions. Our key insight is that by building semantic understanding in MR, we can develop a system that not only greatly enhances user experience through object-specific behaviors, but also it paves the way for solving complex interaction design challenges. In this paper, our proposed framework generates semantic properties of the real-world environment through a dense scene reconstruction and deep image understanding scheme. We demonstrate our approach by developing a material-aware prototype system for context-aware physical interactions between the real and virtual objects. Quantitative and qualitative evaluation results show that the framework delivers accurate and consistent semantic information in an interactive MR environment, providing effective real-time semantic level interactions.

CCS Concepts

•***Computing methodologies*** → *Scene understanding;* •***Human-centered computing*** → *Interaction design;*

1. Introduction

Mixed Reality (MR) as a cross-cutting technology combines computer vision with information science and computer graphics. It seamlessly connects a virtual space with the real world by not only superimposing computer-generated information onto the real-world environment, but also creating new user interactions for novel user experience. Although the boundary definition between MR and Augmented Reality (AR) is blurry. In this paper, we define AR as simple overlays of virtual graphics or information onto images; whereas MR is a general environment which users can interact with and manipulate both physical and virtual items as well as the environment [Int]. This interactive technology will soon become ubiquitous in many applications, ranging from personal information systems, industrial and military simulations, office uses, digital games to education and training.

† Corresponding Author

The latest research in Simultaneous Localisation and Mapping (SLAM) with greater camera tracking accuracy and robustness has enabled the rapid development of Mixed Reality. Although sparse SLAM systems [DRMS07] [KM07] [MAMT15] are proven to be efficient in 3D tracking for monocular cameras, structural information is still missing from the system. In contrast, dense SLAM algorithms [NLD11] [NIH*11] [NFS15] have enabled the construction of object surfaces in order to provide geometric information of the real scene and allow geometric-based interactions in MR environments. Collision effects between virtual and real-world objects in these geometry-aware MR systems increase the immersion of user experience (as can be seen in Figure 2 (a) and (b) for the Ball Pit MR game in Microsoft HoloLens). However, since the individual semantic property of various real-world objects remains undetected, geometry-aware MR systems are unable to distinguish object behaviours due to different object properties. Such shortcomings could always generate uniform object interactions. The lack of higher-level context awareness regarding object properties may break the immersion continuum of users of such systems [GLZR17].

A natural first step moving away from purely geometric-based approaches is to enable semantic understanding of the real environment within MR, hence, generating context-aware interactions. Semantic segmentation [GGOEO*] [SLD17] [ZJRP*15] [CPK*17] [BKC17] leading to greater understanding of the environment is not new to computer vision. However, there is few reported prior work that utilizes the semantic information in MR. Semantic-based interactions via object understanding in MR presents a number of additional challenges: (1) Most of semantic segmentation approaches cannot be running in real-time for interactive environments; (2) It is hard to associate semantics with the structural information of the environment seamlessly on-the-go; (3) Retrieving semantics then generating appropriate interactions is difficult for high computational performance and accuracy.

Realistic interactions in MR require not only geometric and structural information but also semantic understandings of the scene. Embedding semantic information extracted from a 2D image space into the 3D structure of an MR environment is hard, because of the required high accuracy from the semantic segmentation. Careful considerations are needed when designing semantic-based MR interactions. While geometric structure allows accurate object localization, augmentations, and placements, at the user experience level, semantic knowledge is the key to enable realistic interactions between the virtual and real objects. Realistic physical interactions (e.g. a virtual glass shunted on a real concrete floor in an MR environment) can be achieved through the semantic scene understanding. More importantly, using semantic scene descriptions, we can develop high-level tools for efficient design and constructions of large and complex MR applications.

In this paper, we propose a novel context-aware semantic MR framework to address the research gap. Our proposed framework is a 2D-to-3D-to-2D computational pipeline. We demonstrate its effectiveness through example applications. To generate context-aware interactions, we use an end-to-end Deep Learning (DL) framework and a dense Simultaneous Localisation and Mapping (SLAM) algorithm for semantic information integration in MR environment.

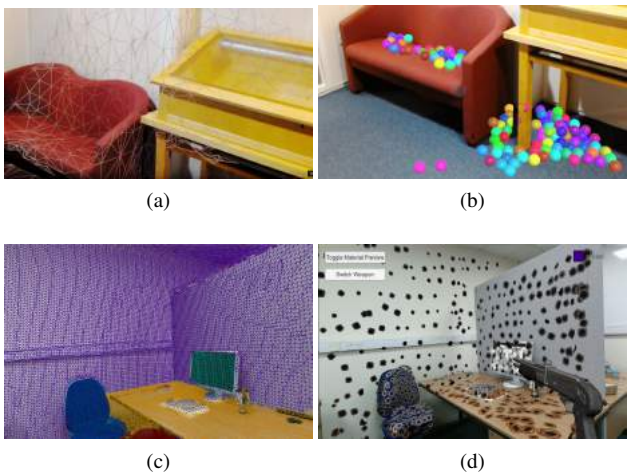


Figure 2: (a): Microsoft HoloLens is capable of reconstructing a MR environment by its built-in "spatial mapping" function and provide a geometric mesh for geometry-based interaction. (b): The Ball Pit MR game based on geometry interactions for Microsoft HoloLens. (c) Our proposed framework can provide a semantic mesh for more advanced context-aware interactions. (d) A shooting game based on our proposed framework. Note that the bullet holes on different objects are different according to material properties of the objects.

We present the labeling of material properties of the real environment in 3D space as a novel example application to deliver realistic physical interactions between the virtual and real objects in MR. To the best of our knowledge, this is the first work that presents a context-aware MR by using deep learning based semantic scene understanding and generating semantic interactions at the object-specific level. Our approach is one step further towards the high-level conceptual interaction modelling in complex MR environment for enhanced user experience.

Dense SLAM KinectFusion [NIH*11] was used for camera pose recovery and 3D model reconstructions to create a classic geometry-aware MR environment first. We trained Conditional Random Fields as Recurrent Neural Networks (CRF-RNN) [ZJRP*15] by using a large-scale material database [BUSB15] for detecting material properties of each object in the scene. The 3D geometry/model of the scene is then labeled with the semantic information of materials that made up of the scene. Therefore, with the semantic information available, realistic physical interactions between objects can be generated based on the material properties of the objects which the user is interacting with. Figures 2 (c) and (d) show a shooting game example for object-specific material-aware interactions. Our framework is based on a 3D volume semantic construction, classification and occupancy mapping for the real objects (i.e.2D-to-3D-to-2D computational pipeline.

Our proposed framework is both efficient and accurate in seman-

tic labeling and inference for generating realistic context-aware interactions. Two tests are designed to evaluate the effectiveness of the framework. One test evaluates the end-to-end system accuracy by comparing the dense semantic ray-casting results with manually labeled ground truth of 25 key-frames of two different scenes. Another test is a user experiment with 28 participants to qualitatively evaluate user experience under three different MR conditions. The results show that the framework delivers more accurate 3D semantic mappings than directly using the state-of-the-art 2D semantic segmentation.

In the next section, we review related work on geometry-based MR Interactions and recent approaches to semantic segmentation using Convolutional Neural Network (CNN). The following sections introduce our framework with SLAM dense reconstructions of the scene and the 3D semantic fusion, and describe our implementation and the evaluation framework. Finally, we demonstrate our results compared with the state-of-the-art semantic segmentation algorithms.

2. Previous Work

Our approach draws upon the recent success of dense SLAM algorithms [NLD11] [NIH*11] [NFS15] and deep learning methods for semantic segmentations [GGOEO*] [SLD17] [ZJRP*15] [CPK*17] [BKC17] that have been mostly used in the field of robotics until now.

2.1. Geometry-based MR Interactions

Interaction modeling between virtual and real objects in MR are mostly geometry-based through plane feature detections or full 3D reconstructions of the real world. Methods of using plane detection [SMGKD14] [NOBW16] estimate planar surfaces in the real-world, which virtual objects are placed onto and collided with. Random Sample Consensus (RANSAC) algorithm [FB81] estimates planar surfaces based on sparse 3D feature points extracted from a monocular camera. Plan detection does not require a depth camera, is computationally efficient and can run on mobile phones, which is shown in the newly released Mobile AR systems [App17] [Goo17]. One obvious shortcoming of the plane detection is the requirement for large planar surfaces to delivery MR interactions. Collision meshes for non-planar surfaces are impossible, hence, restricting user to only interact in certain areas and with types of objects.

Recent advances in depth sensors, display technologies and SLAM software [NLD11] [NIH*11] [WJK*13] [NFS15] have opened up the potential of MR systems. Spatial structures of the real environment can be generated at ease to provide accurate geometries for detecting collisions between virtual and real objects. Examples of geometry-based interactions include: a virtual car 'drives' on an uneven real desk [NLD11]; the Super Mario game played on real building blocks [KTY*13], and the Ball Pit game in HoloLens [Mic17], where Figure 2 (a) and (b) illustrate the concept. Impressive as they are, the state-of-the-art systems are still limited to basic and uniform geometry-based virtual and real object interactions. Lacking high-level semantic descriptions and basic scene understandings in MR have compromised the immersion of user experience, and the realism of interactions is reduced and easily broken. One example is in the Ball Pit game, material properties of the real objects are not recognized. Thus, a ball falling onto a soft surface would still bounce back unrealistically against the law of physics.

2.2. Deep Semantic Understanding

Semantic segmentation is an emerging technology in computer vision. The recent success of CNN has achieved semantic level image recognition and classifications with great accuracy [KSH12], enabling many novel applications. In the last few years, more complex neural networks such as FCN [SLD17], CRF-RNN [ZJRP*15], DeepLab [CPK*17] and SegNet [BKC17] have enabled image understanding at the pixel level. After being trained on large-scale databases, these networks can predict and label semantic information at every pixel of an image.

Recently, the joint learning of depth and semantics [PXZ*15] [JCSL18] [ZCX*18] approaches have achieved better results than a single learning task, however, the joint learning means a higher demand of data, so the availability of datasets is a challenge. With the development of 3D CNN, [QSMG17] [LS18] achieved novel learning of semantic information from 3D point clouds. However, the prediction of semantic attributes only in 3D does have its limitations – for example, for the material-aware interaction proposed in our paper, the material information is very hard to be inferred from 3D level (a chair-shape object can be made of wood, but can also be plastic or metal). Most of the semantic information about 3D point clouds is regarding the 3D shapes and geometry. In contrast, 2D semantic segmentation algorithms have been proven efficient and effective.

Combined with SLAM systems, 2D semantic segmentation can be achieved in 3D environments [RA17] [TTLN17] [ZSS17] [MHDL17], a promising future in robotic vision understanding and autonomous driving. Unlike these existing methods that aimed at providing the semantic understanding of the scene for robots, we are focusing our attention on human interactions. Our goal is to provide the user with realistic semantic level interactions in MR. In this paper, we use MR as a bridge to connect AI and human for a better understanding of the world via intelligent context-aware interactions.

2.3. Context and Semantic Awareness in MR Environment

Prior approaches have studied context and semantic understandings in 3D virtual environments, for example, semantic inferring in interactive visual data exploration [NEF12]; enhancing software quality for multi-modal Virtual Reality (VR) systems [FWL17]; visual text analytics [EFN12]; and interactive urban visualization [DZMQ16]. Context awareness is also introduced in computer-aided graphic design such as inbetweening of animation [Yan18]; 3D particle clouds selection [YEII16], and illustrative volume rendering [RBG07]. Virtual object classifications are proposed in VR applications using semantic associations to describe virtual object behaviors [CTB*12]. The notion of *conceptual modeling* for VR applications is pointed out by Troyer *et al.*, highlighting a large gap between the conceptual modeling and VR implementations. It is

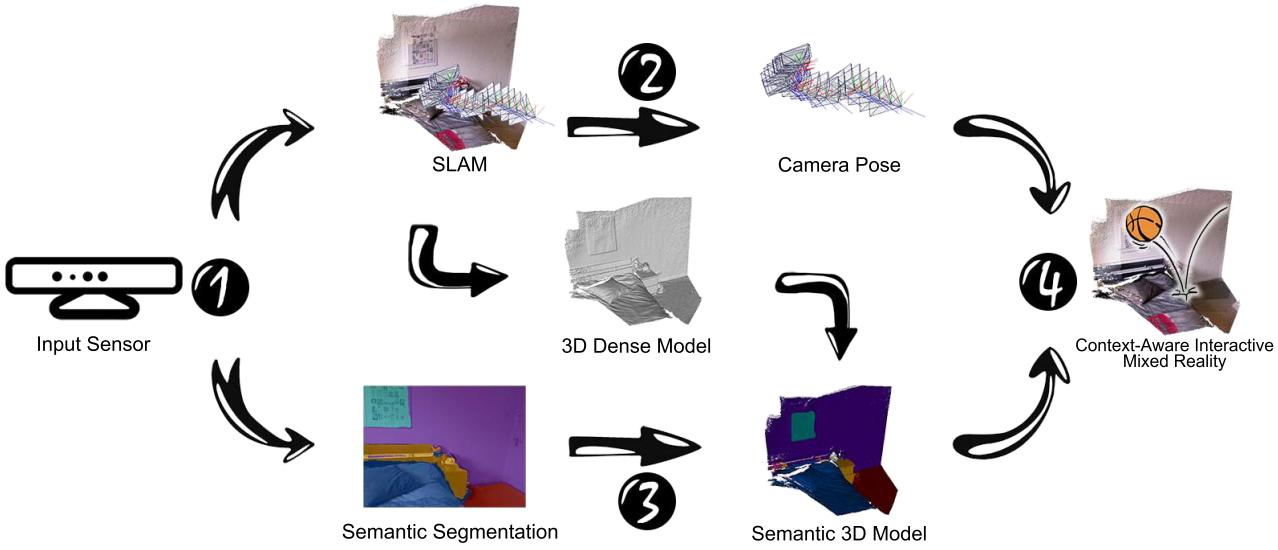


Figure 3: Flowchart demonstrates the whole framework. Starting from (1) an input sensor, frames go through (2) Tracking and Reconstruction stream and (3) Semantic Segmentation and Fusion stream, and finally the poses and semantic 3D model are used to provide the Context-Aware Interactive Mixed Reality.

suggested taking a phased approach (i.e. conceptual specification, mapping, and generation phases) to bridge the gap [DTKPB07].

Recently, the idea of extending Augmented Reality (AR) applications to become context-aware has been presented in computer graphics [GLZR17], which proposes to classify context sources and context targets for continuous user experience. A method is proposed for authentic simulating outdoor shadows to achieve seamless context-aware integration between the virtual and real objects for mobile AR [BBBM18].

2.4. The contributions of our work

Although SLAM and semantic segmentation are active research topics in the robotics and computer vision communities, there is a lack of work in computer graphics research by combining these approaches together for higher level context-aware MR interactions. In this paper, we address the idea of "ubiquitous interaction" in the MR environment to explore a deep semantic understanding of the environment and take a step further towards the high-level interaction design for MR. The contributions of this paper are:

- We propose the concept of "Context-Aware Mixed Reality" in which everything in the real-world is interactable in the virtual MR environment.
- We present a general framework for the proposed "Context-Aware Mixed Reality" and implement two MR games to demonstrate the concept.
- We propose a fusion method with the 3D geometric label optimization for improving the accuracy of 3D semantic mesh.
- We have conducted the accuracy study and user experience evaluations for the quantitative evaluation of the system accuracy and the improvement to immersive and natural interactions.

Empowered by the latest computer vision and AI technologies, our proposed framework can intelligently assist the interaction design in MR and seamlessly generates the higher level virtual-real interactions to increase the realism and naturalism of interactions in MR and deliver immersive user experience greatly.

3. Framework Overview

Figure 3 shows the proposed framework. Starting from an ①***Input Sensor***, two main computation streams are constructed: ②***Tracking & Reconstruction Stream*** and ③***Context Detection & Fusion Stream***, which are finally merged and output to the ④***Interactive MR Interface*** for generating context-aware virtual-real interactions.

3.1. Input Sensor

An *input sensor*, an RGB-D camera such as Microsoft Kinect, ASUS Xtion series or built-in sensors on Microsoft HoloLens, is used to acquire the depth information directly for the 3D reconstruction of the environment. Monocular or stereo cameras would also work if combined with dense SLAM systems [NLD11], but the accuracy and real-time performance of Mono devices are not guaranteed.

3.2. Camera Tracking & Reconstruction Stream

The *tracking & reconstruction stream* shown in the upper path of Figure 3 processes the video captured by the input sensor. A SLAM system continuously estimates the camera pose and simultaneously reconstruct a 3D dense model. This is a typical method used in the latest MR systems such as Microsoft HoloLens for implementing

geometry-aware MR. A dense 3D model serves as a spatial collision mesh and the inverse of the camera pose extracted from the SLAM guides the movement of the collision mesh to visually correspond to the real-world objects.

3.3. Context Detection & Fusion Stream

The lower path of Figure 3 shows the *Context Detection Stream*. Image sequences from the input sensor are context sources to be processed by semantic segmentation algorithms to output dense pixel-wise object attributes and properties of the scene. Based on the semantically segmented information, The context information relevant to implementing context-aware experience is generated. The 2D semantic segmentation results are then projected onto the scene and fused with the 3D dense model (from *tracking & reconstruction stream*) to generate a semantic 3D model based on the camera pose of the current frame.

3.4. Interactive MR Interface

The semantic 3D model is combined with the camera pose to provide a context-aware MR environment. High-level interactions are designed based on the semantics , and tools can be developed to facilitate the design and the automatic construction of complex MR interactions in different applications.

The advantages of the proposed framework are:

1) **Accurate 3D Semantic Labeling**: The Context Detection & Fusion Stream can predict a pixel-wise segmentation of the current frame, which is further fused onto the 3D dense model. The system builds a semantic 3D model that contains voxels, each voxel encoding the contextual knowledge of the environment. The voxel-based context-aware model delivers the semantic information through ray-cast queries about the object properties to generate different user interactions. Object properties can be high-level descriptions (e.g.types of materials and interaction attributes.)

2) **Real-time Performance**: In deep learning based approaches, the semantic segmentation is computationally expensive especially when processing frame by frame in real-time applications. The proposed framework achieves the real-time performance by embedding the semantic information into the 3D dense model after the initial segmentation process, and can be updated along with time. In doing so, the semantic segmentation do not need to be processed at each frame, but the whole system can deliver thereal-time semantic information.

3) **Computer-Aided Interaction Design**: With the context information available, virtual and real object interactions can be designed and computed by feeding attributes of the real world objects to the target software module for processing (a physics module or an agent AI module). For example, realistic physical interactions between the virtual and real objects can be computed by feeding the material properties of the real world objects to physically-based simulation algorithm (such as our throwing plates game in the following section).

4. Implementation

We present our novel MR framework in the context of object material-aware interactions as an implementation example to demonstrate the concept of context-aware MR. Material properties of the real world in a MR environment enable the generation of realistic physical interactions. The example implementation is also used in the accuracy study and the user experiment presented in the following sections.

4.1. Camera Tracking and Model Reconstruction

A dense SLAM system [NIH*11] is used to achieve the accurate camera tracking and dense 3D model reconstructions of the environment, which estimates camera poses and reconstructs the 3D model in real-time. Depth images from a Kinect sensor are projected into the 3D model. The camera pose and a single global surface model can be obtained simultaneously through a coarse-to-fine iterative closest point (ICP) algorithm. The tracking and reconstruction processes consist of four steps:

1) Each pixel acquired by the depth camera is transformed into the 3D space by the camera's intrinsic parameters, and the corresponding depth value is acquired by the camera;

2) An ICP alignment algorithm is performed to estimate the camera poses between the current frame and the reconstructed global model;

3) With the available camera poses, each consecutive depth frame can be fused incrementally into one single 3D reconstruction by a volumetric truncated signed distance function (TSDF);

4) Finally, a ray-casting process is used to predict a surface model

A Microsoft Kinect is used as the input sensor with an OpenNI2 driver to capture RGB images and calibrated depth images at the resolution of 640 x 480 at 30 frames per second (FPS).

4.2. Deep Learning for Material Recognition

We have trained a deep neural network for the 2D material recognition task. Our neural network is implemented in *caffe* [JSD*14] based on the CRF-RNN architecture [ZJRP*15], which combines the FCN with Conditional Random Fields (CRF) based on the probabilistic graphical modelling for contextual boundary refinement. We have used the Materials in Context Database (MINC) [BUSB15] as the training database that contains 3 million labeled point samples and 7061 labeled material segmentations in 23 different material categories.

The VGG-16 pre-trained model for ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [SZ15] is used as the initial weights of our neural network. Based on the MINC dataset, we then fine-tuned the network from 1000 different classes (ImageNet containing 1000 classes of labels) to 23 class labels as the output. VGG-16 is a CNN model specifically designed for classification tasks and only produces a single classification result for a single image. Therefore, we manually cast the CNN into a Fully Convolutional Network (FCN) for pixel-wise dense outputs [SLD17]. By

transforming the last three inner product layers into convolutional layers, the network can learn to make dense predictions efficiently at the pixel level for tasks such as semantic segmentation. The fully-connected CRF model is then integrated into FCN to improve the semantic labeling results.

Fully-connected CRF encodes pixel labels as random variables to form a Markov Random Field (MRF) [KS80] conditioned on a global observation (the original image). By minimizing the CRF energy function in the Gibbs distribution [LRKT09], we obtain the most probable label assignment for each pixel in an image. With this process, the CRF refines the predicted label using the contextual information. It is also able to refine weak semantic label predictions to produce sharp boundaries and better segmentation results (see Figure 10 for the comparison of FCN and CRF-RNN). During the training process, CRF is implemented by multiple iterations, each takes parameters estimated from the previous iteration, which can be treated as a Recurrent Neural Network (RNN) structure [ZJRP*15].

As the error of CRF-RNN can be passed through the whole network during a backward propagation, the FCN can generate better estimations for the CRF-RNN optimization process during the forward propagation. Meanwhile, CRF parameters, such as weights of the label compatibility function and Gaussian kernels can be learned from the end-to-end training process.

We use 80% of the 7061 densely labeled material segmentations in the MINC dataset as the training dataset and the rest of 20% as testing sets. The training dataset is trained using a single Nvidia Titan X GPU for 50 epochs, after which there is no significant decrease in loss. For testing results, we obtain a mean accuracy of 78.3% for the trained neural network. The trained network runs at around 5 frames per second for the 2D dense semantic segmentation at the resolution of 480x270 pixels on a Nvidia Titan X GPU. We input 1 frame into our neural network for every 12 frames according to our test to achieve a trade-off between the speed and accuracy.

4.3. Bayesian Fusion for 3D Semantic Label Fusion

The trained neural network for material recognition only infers object material properties in the image space. As the camera pose for each image frame is known, we can project the semantic labels onto the 3D model as textures. A direct mapping can cause information overlapping, since accumulated weak predictions and noises can lead to a bad fusion result as shown in Figure 5 (a), where boundaries between different materials are blurred. We solve this issue by utilizing the dense pixel-wise semantic probability distribution produced by the neural network over every class. Therefore, we can improve the fusion accuracy by projecting the labels with a statistical approach using the Bayesian fusion [ASZ*16] [HFL14] [ZSS17] [MHDL17]. Bayesian fusion enables us to update the label prediction l_i on 2D images I_k within the common coordinate frame of the 3D model.

$$P\left(x=l_i|I_{1,\ldots,k}\right)=\frac{1}{Z}P\left(x=l_i|I_{1,\ldots,k-1}\right)P\left(x=l_i|I_k\right) \quad (1)$$

where Z is a constant for the distribution normalization. The label of each voxel is updated with the corresponding maximum probability $p\left(x_{max}=l_i|I_{1,\ldots,k}\right)$. The Bayesian fusion guides the label fusion process and ensures an accurate mapping result over time to overcome the accumulated errors to some extent. Figure 5 (a) shows without the Bayesian fusion, the label fusion results are less clear due to the overlapping of weak predictions. In contrast, 5 (b) with the Bayesian fusion, the fusion results are much cleaner.

After semantic information fusion into the 3D model, we can get a semantic labeled 3D model. Although the Bayesian fusion is used to guide the fusion process, due to the accumulation of the 2D segmentation error and the tracking error, in some area, the semantic information still not perfectly matches the model structure (see Figure 4). Next, we explain how to further improve the fusion accuracy by proposing a new CRF label refinement process on 3D structures.

4.4. 3D Geometric Label Optimization

We further improve the accuracy of the 3D labelling with a final refinement step on the semantic information using the structural and color information of vertices of the 3D semantic model. From the fully connected CRF model, the energy of a label assignment x can be represented as the sum of unary potentials and pairwise potentials over all i pixels:

$$E\left(x\right)=\sum_i \psi_u\left(x_i\right)+\sum_i\sum_{j\in N_i}\psi_p\left(x_i,x_j\right) \quad (2)$$

where the unary potential $\psi_u\left(x_i\right)$ is the cost (inverse likelihood) of the i_{th} vertex assigning with the label x. In our model implementation, we use the final probability distribution from the previous Bayesian Fusion step as the unary potential for each label of every vertex. The pairwise potential is the energy term of assigning the label x to both i_{th} and j_{th} vertices. We follow the efficient pairwise edge potentials in [KK11] by defining the pairwise energy term as a linear combination of Gaussian kernels:

$$\psi_p\left(x_i,x_j\right)=\mu\left(x_i,x_j\right)\sum_{m=1}^{M}w^{(m)}k_G^{(m)}\left(f_i,f_j\right) \quad (3)$$

where w^m are the weights for different linear combinations, k_m^G are m different Gaussian kernels that f_i and f_j correspond to different feature vectors. Here, besides the commonly used feature space in [KK11] [ZJRP*15] such as the color and the spatial location, the normal direction is also considered as a feature vector to take the full advantage of our 3D reconstruction step:

$$\begin{aligned}k_G\left(f_i,f_j\right)=&w^{(1)}exp\left(-\frac{\left|p_i-p_j\right|^2}{2\theta_p^2}\right)\\&+w^{(2)}exp\left(-\frac{\left|p_i-p_j\right|^2}{2\theta_{pI}^2}-\frac{\left|I_i-I_j\right|^2}{2\theta_I^2}\right)\\&+w^{(3)}exp\left(-\frac{\left|p_i-p_j\right|^2}{2\theta_{pn}^2}-\frac{\left|n_i-n_j\right|^2}{2\theta_n^2}\right)\end{aligned}$$

where p_i and p_j are pairwise position vectors; I_i and I_j are pairwise RGB color vectors; n_i and n_j are pairwise normal directional vectors. The first term is the smoothness kernel assuming that the nearby vertices are more likely to be in the same label, which can

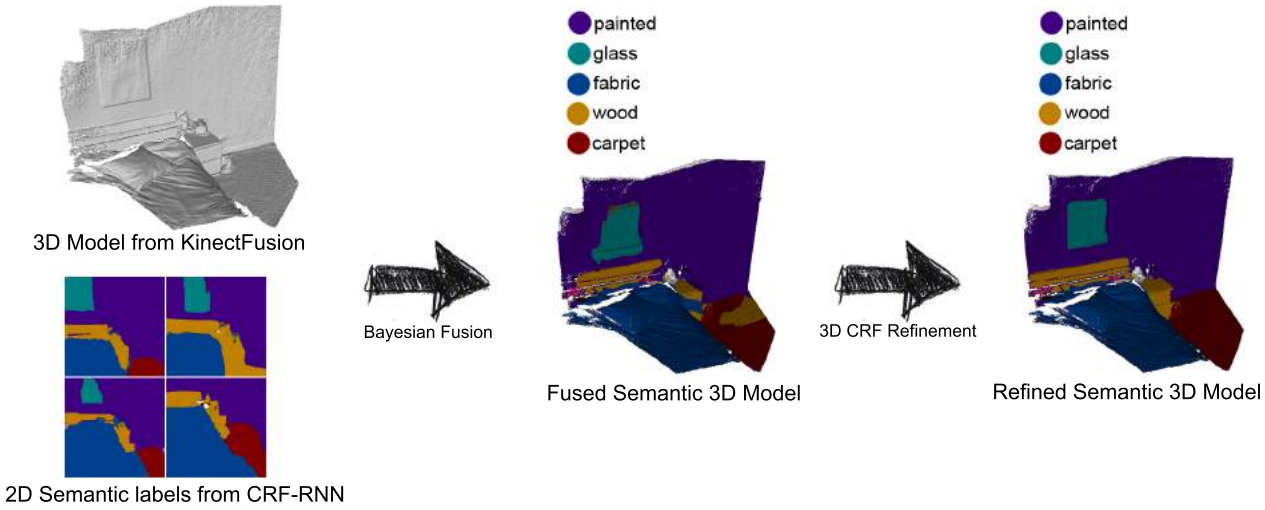
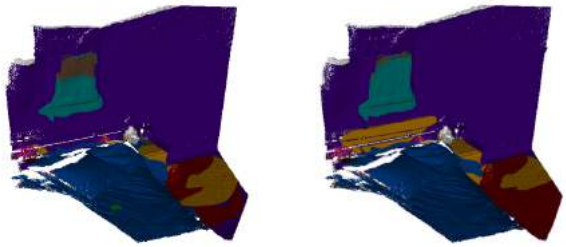


Figure 4: 3D semantic label fusion and refinement. From left to right: the 3D model and 2D semantic segmentation results are fused by Bayesian Fusion into semantic 3D model; the semantic 3D model goes through the 3D CRF refinement for improving quality.



(a) (b)

Figure 5: (a) 3D semantic label fusion using direct mapping. (b) 3D Semantic Label Fusion with Bayesian fusion.

efficiently remove small isolated regions [SWRC09] [KK11]. The second term represents the appearance kernel that takes into account of color consistency, since the adjacent vertices with similar color(s) are more likely to have the same label. The third term is the surface kernel which utilizes the 3D surface normal as a feature that vertices with similar normal directions are more likely to be the same label.

By minimizing Equation 2, semantic labels on our 3D model are further refined according to the colour and the geometric information, which can efficiently eliminates the ”label leaking” problem caused by the 2D semantic segmentation errors and the camera tracking errors (see Figure 4).

4.5. Interaction Interface

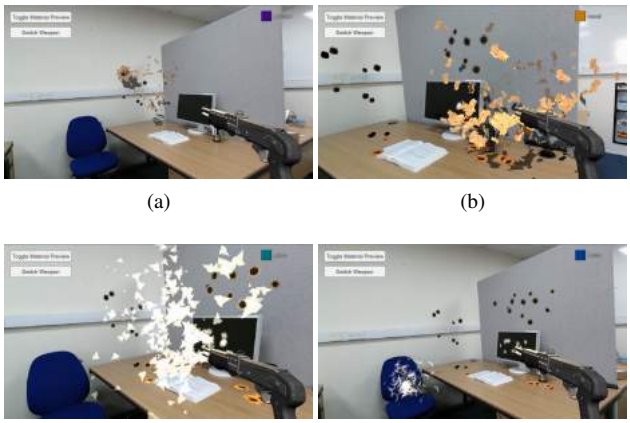
A user interface is developed with two layers. The top layer displays the current video stream from a RGB-D camera, whilst the semantic 3D model serves as a hidden physical interaction layer to provide the interactive interface. In the interactive MR application, a virtual camera is synchronized with predicted camera poses for projecting the 3D semantic model onto the corresponding point of view of the video stream. Figure 9 shows that the back layer of the interface displays the video stream from a RGB-D camera; A semantic interaction 3D model is in the front of the video layer for handling interactions of different materials (Green: glass, Purple: painted, Blue: fabric, Yellow: wood, Red: carpet). The virtual and the real physical interactions are performed on the interaction model. The context-aware interaction model is invisible to allow users to interact with the real-world objects to experience an immersive MR environment. The interaction layer also computes real-time shadows to make the MR experience even more realistic. An oct-tree data structure accelerates the ray-casting queries for the material properties to improve the real-time performance. Finally, corresponding physical interactions based on the semantic information of different materials are achieved through physics simulations.

5. Example Applications

5.1. Context-Aware Interactive Games

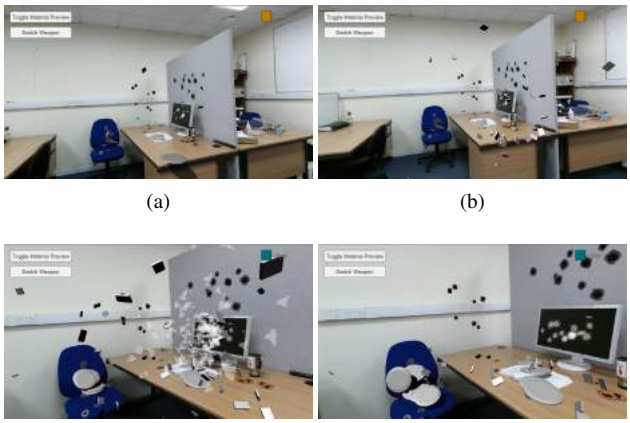
Based on our context-aware MR framework, two FPS games are developed to demonstrate the concept of the proposed material-aware interactive MR. Next, we describe the design of interactions and evaluations.

Games are interaction demanding applications that are driven by computational performance and accurate interactions in virtual spaces. We have designed two MR games that can directly interact with the real-world objects. A shooting game is designed to show material-aware interactions between bullets and the real world objects. The shooting scenario is chosen because we want to test the accuracy of the semantic 3D model using ray-cast queries. In this



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

Figure 6: The screenshots of our MR shooting game. The interaction is different when shooting (a)walls, (b)desks, (c)computer screen and (d)chair.



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

Figure 7: The screen shots of our MR throwing plates game. The interaction is different when throwing plates to (a) book, (b) desks, (c) computer screen, and (d) chair.

game, as shown in Figure 6, multiple interactions for different materials have been implemented including different bullet holes, flying chips and hitting sound when hitting different objects: (a)walls, (b)desks, (c)computer screen, and (d)chair. The interaction for different material context is as real as possible.

Another way to show the capability of the context-aware framework is to match the interaction results to the user's anticipation of the interaction results using everyday scenarios that familiar to users, testing the immersive experience of the MR system from the user's perspective. The second example is designed to match user expectations for material-specific physical interactions.

As shown in Figure 7, users throw virtual plates onto real-world objects of the MR environment, resulting in material-aware physical interactions induced by various material properties of the real objects. In Figures 7 (a) and (b), virtual plates are broken when falling onto the desk, but bounced back when colliding with a book; in (c) when colliding with a computer screen, the plate is broken with the flying glass chips; in (d), the plate remains intact colliding with a soft chair.

5.2. Context-Aware Relighting

When virtual illuminated objects are involved in MR environment, relighting is an important technology that makes the user experience more realistic and immersive. With the 3D model available, it is possible to re-render the scene (such as Figure 8 (a)) but lack of realism due to the unknown lighting property and constant reflectance.

Based on our proposed system, high fidelity 3D models can be acquired from SLAM reconstruction and the semantic mapping from semantic segmentation can provide the high-level context of object properties (such as material, metallicity, smoothness, etc.). Then the material-specific relighting can be done automatically through the render engine (Unity3D in our example), by determining the specular and diffuse reflections based on the object property (as shown in Figure 8 (b), that the screen was set to specular reflections while other materials were diffuse reflections). And this material-specific relighting is more realistic and accurate than previous relight work [ZCC16] that relays on the Manhattan-world assumptions [CY00] without any object-level relighting.

6. Experimentation

6.1. Experimentation setup

The RGB and calibrated depth image sequences were acquired from Microsoft Kinect at ~30 FPS at the resolution of 640 x 480. The semantic segmentation neural networks, SLAM and semantic fusion modules are running as back-end services to provide semantic mesh for front-end game interface through Unity3D.

For the back-end services, our semantic segmentation networks were trained and running in Tensorflow on a workstation with a single Nvidia Titan X GPU (12G Memory), which can be running at ~3 FPS at the resolution of 640 x 480. The SLAM tracking and dense surface reconstruction module can be running at ~20 FPS.

For the front-end game interface, the background camera and virtual objects were rendered at ~24 FPS, while the semantic mesh was updated at ~1 FPS.

6.2. Accuracy Study

Multiple factors affect the accuracy of the system: (1) the camera tracking, (2) the 3D model reconstruction; (3) the deep semantic material segmentation; (4) the 2D to 3D semantic model fusion; and (5) the implementation of the ray-casting. As the goal of our proposed framework is to deliver real-time and accurate semantic interaction in MR, so we would like to evaluate the dense ray-casting queries of the 3D semantic model, and we transform the dense ray-casting queries into a 2D projection from our semantic 3D model

(a) Uniform relighting with constant reflectance

(b) Material-specific relighting by our proposed framework

Figure 8: A virtual torch in MR to illustrate the relighting in MR.

for evaluation. Here, we not only evaluated the whole framework (3D-2D w/ CRF in Table 1), but we also conducted an ablation accuracy study to demonstrate how each component contributed to the final result. A total of 25 key-frames from two different scenes (office and bedroom) are selected, and at the same time, the 2D projections of the 3D semantic models are captured as the dense ray-casting query results at the corresponding key-frames (see Figure 11). Ground truth for the accuracy evaluation is obtained by manually labelling 25 RGB images with the same material labels. The four common evaluation criteria [SLD17] [ZJRP*15] for semantic segmentation and scene parsing evaluations are used to evaluate the variations of pixel accuracy and region intersection over union (IoU).

1. pixel accuracy $\frac{\sum_i n_{ii}}{\sum_i t_i}$
2. mean accuracy $\frac{1}{n_{c1}} \sum_i \frac{n_{ii}}{t_i}$
3. mean IoU $\frac{1}{n_{c1}} \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}$
4. frequency weighted IoU $\frac{1}{\sum_k t_k} \sum_i \frac{t_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}$

where n_{ij} represents the the number of pixels of class i predicted to be class j; n_{c1} is the total number of classes; $t_i = \sum_j n_{ij}$ is the total number of pixels of class i.

Table 1: Accuracy study results

Stage	pixel acc.	mean acc.	mean IoU	f.w. IoU
FCN [SLD17]	81.61	63.69	49.54	76.16
CRFRNN [ZJRP*15]	85.68	51.73	41.32	79.76
3D-2D w/o CRF	**87.86**	**70.69**	**54.81**	**81.86**
3D-2D w/ CRF	**89.42**	**72.06**	**56.32**	**84.30**

As can be seen from Table 1, after 2D-3D fusion, 3D CRF refinement and finally 3D-2D projections, our framework can provide more accurate semantic segmentation results compared with the 2D methods such as FCN and CRF-RNN. Figure 10 shows some semantic segmentation samples. Taking the advantages of the 3D constraints and refinement in our proposed framework, our semantic segmentation results are more uniform, sharp and accurate.

6.3. User Experience Evaluation

We conducted an interactive user study to evaluate the effectiveness of the semantic-based MR system. Using the throwing plates game, three test conditions are designed by setting different collision responses:

1) *No Collision Mesh*: Virtual plates were thrown into the real world without any collision being detected. The plates will fly to infinite distance.

2) *Uniform Collision Mesh*: Virtual plates interact with the real world with the uniform collision mesh being activated, but no object-specific interaction is generated. The plates will break when being contacted with any object

3) *Semantic Collision Mesh*: Physics responses of the virtual plates with the real-world objects are dependent on the material properties of the objects in the real world. The plates will break when contacted with defined hard objects: desk, wood; The plates will not break when contacted with defined soft objects: fabric

The objective of the user study is to assess the realism of the MR environment by measuring how much the semantic-based interaction matches the user anticipation. We investigate whether or not the semantic-based interactions can significantly improve the realism of MR systems and delivers immersive user experience.

Firstly, we evaluate the realism of physical interactions such as collision responses in MR systems. We test to see if users are able to detect differences in these three interaction conditions between virtual and real objects, and whether or not the realism in MR can be improved via context-aware physical responses. Secondly, to ensure the quality of qualitative study, we test if there is any risk that the user experience of the proposed MR system could be affected by gender factor and their previous engagement with MR or VR technologies.

6.3.1. Participants

We recruited 28 undergraduate students (22 were identified as

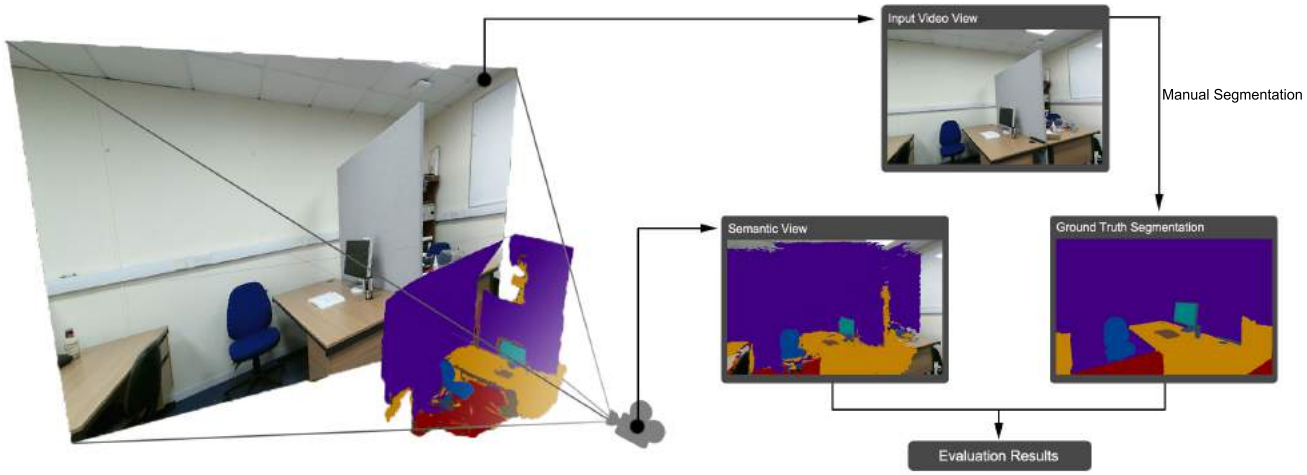


Figure 9: The evaluation framework

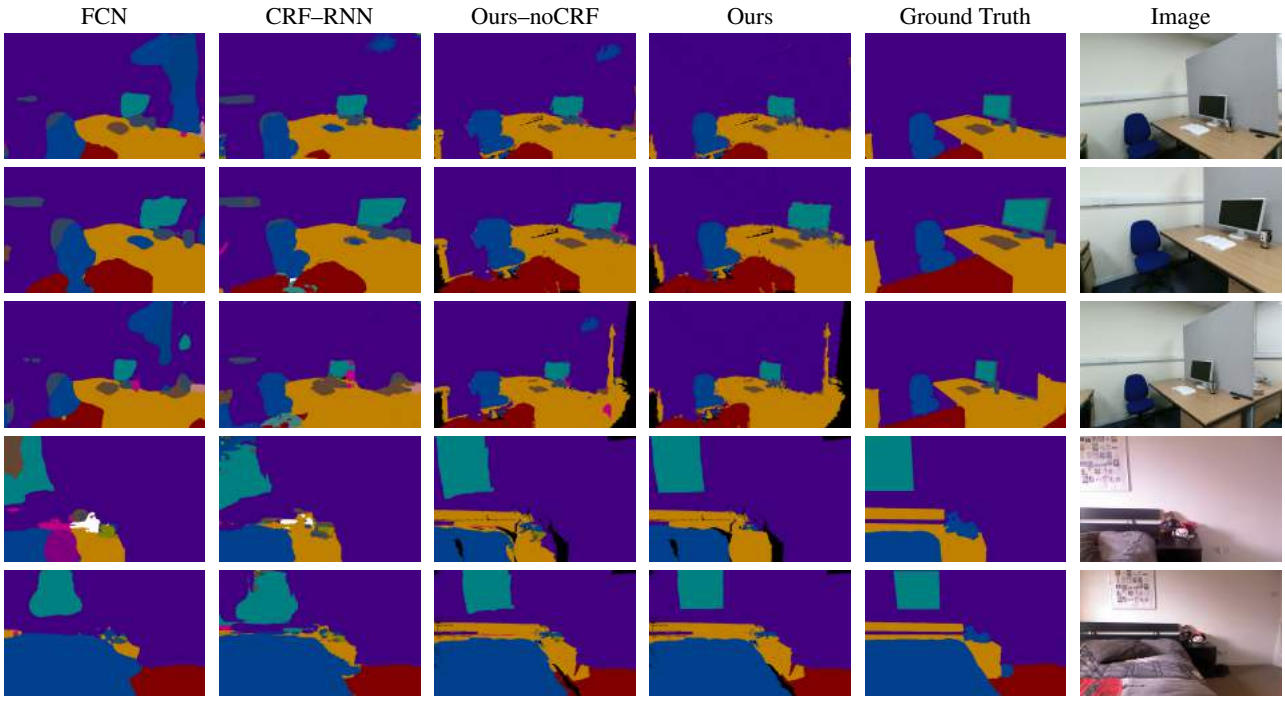


Figure 10: Semantic segmentation samples for each column from left to right: FCN, CRF-RNN, Final results without CRF refinement, Final results with CRF refinement, Ground Truth, Input image

males, 6 as females) for our user experience experiment. The session included a game play part and a questionnaire part. Each participant was invited to a room with a computer to play the "throwing plates" MR game with 3 pre-defined conditions. On the screen there was a simple interface with 3 buttons: "Game 1", "Game 2" and "Game 3" for the "throwing plates" MR game with "No Collision Mesh", "Uniform Collision Mesh", and "Semantic Collision Mesh" conditions. The users can select to enter which game, and can replay the MR game with each condition as much as they want, so that the participant can take time to digest and answer the questions. After 3 conditions were well played, the participants were asked to fill a questionnaire with 4 questions: the first question is to ask whether the participant had any previous experience with VR/AR games, the other 3 questions are the rating of the MR game experiences on the scale of 1 (very bad) to 10 (very good) based on the quality of the MR interactions and realism.

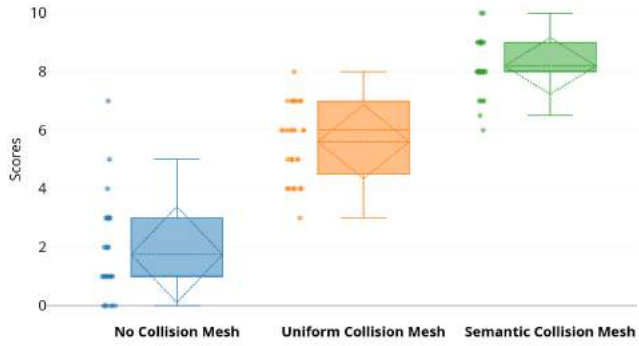


Figure 11: The box plot of the user experience evaluation results of the three conditions: Blue: No Collision Mesh; Orange: Uniform Collision Mesh; Green: Semantic Collision Mesh

Table 2: The user experience evaluation results of three conditions.

Groups	N	Mean	Std. Dev.	Std. Err.
No Collision Mesh	28	1.75	1.6694	0.3155
Uniform Collision Mesh	28	5.6071	1.2864	0.2431
Semantic Collision Mesh	28	8.1964	0.9845	0.1861

6.3.2. Results

We used the score from 1 to 10 as the interval data so that we can use parametric ANOVA to analyse the data. We have performed a repeated measure ANOVA test to analyze scores obtained for the three conditions.

Within-Subjects Results
Mauchly's test shown that the assumption of sphericity is not significant ($X^2 = 2.478, p = 0.290, \varepsilon = 0.907$). So we can reject the hypothesis that the variances of the differences between the three conditions were significantly different. Main effects were found within the three conditions ($F_{2,48} = 152.043, p < 0.001$).

The following post hoc Bonferroni pairwise comparisons show that the *Semantic Collision Mesh* ($M = 8.196, SD = 0.984$) is significantly better than the other two MR conditions ($p < 0.001$), indicating that the proposed semantic interactions through the inference of material properties can greatly improve the realism of the MR system. We also found that the *Uniform Collision Mesh* ($M = 5.607, SD = 1.286$) offers much better MR experience ($p < 0.001$) than the *No Collision Mesh* ($M = 1.750, SD = 1.669$) but less realistic compared with the semantic context-aware MR. The mean scores of the three system conditions are shown in Table 2 and the box plot is shown in Figure 11.

Between-Subjects Results
Furthermore, as our test group has unbalanced gender issue (78.57% are male), and some candidates had ever played VR/AR games for at least once (57.14%). We also conducted a between-subjects repeated measure ANOVA test to reveal whether gender and the previous VR/AR experience have influences on our previous results. It has been shown that the final test results are not affected by gender ($p = 0.210$) or whether the candidate ever played VR/AR games or not ($p = 0.654$).

7. Conclusion and Discussion

We show how deep semantic scene understanding methodology combined with dense 3D scene reconstruction can build high-level context-aware highly interactive MR environment. Recognizing this, we implement a material-aware physical interactive MR environment to effectively demonstrate natural and realistic interactions between the real and the virtual objects. The accuracy study result shows that our 2D-3D-refinement-2D framework can greatly improve the accuracy of delivering the semantic information to users in the context of semantic MR interaction. Our user study reveals that the semantic interaction is very important to the realism of MR experience, and our proposed semantic interaction enabled MR system achieved the best MR user experience in terms of MR interactions. Our work is the first step towards the high-level interaction design in MR. This approach can lead to better system design and evaluation methodologies in this increasingly important technology field.

There are some immediate directions for future research and we mention two such directions now. Although in this paper we focus our discussions on material understanding and its semantic fusion with the virtual scene in MR environment, the concept and the framework presented here are applicable to address many other context-aware interactions in MR, AR and even VR. The framework can be extended by replacing the training dataset with more general object detection databases for constructing different interaction mechanisms and context. Realistic physics-based sound propagation and physics-based rendering using the proposed context-aware framework for MR are promising directions to pursue. Also, our studies are minimal and there are still potential areas for improvements: bigger evaluation dataset, bigger scale of user study, bias-free user study (randomly present the three conditions), design more questions to better understand the whole experience.

Our results have hinted that the study of semantic constructions in MR as a high-level interaction design tool is worth pursuing, as more comprehensive methodologies emerging, complex rich MR applications will be developed in the near future. However it is worth notice that unlike physics, many natural interactions are not easily defined or represented for machines to understand (such as what will happen when pouring virtual water onto a real TV?). We believe that answering such big questions should be an essential part of our future work which will enable the next generation of MR with the help of the next generation artificial general intelligence. We hope that our initiative on the application of AI on MR offers a new insight for the future of MR, and bridges the gap between the virtual and real worlds in context-aware interactions.

References

[App17] APPLE: Arkit - apple developer, 2017. URL: https://developer.apple.com/arkit/. 3

[ASZ*16] ARMENI I., SENER O., ZAMIR A. R., JIANG H., BRILAKIS I., FISCHER M., SAVARESE S.: 3d semantic parsing of large-scale indoor spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016), pp. 1534–1543.

URL: http://dx.doi.org/10.1109/CVPR.2016.170, doi:10.1109/CVPR.2016.170. 6

[BBBM18] BARREIRA J., BESSA M., BARBOSA L., MAGALHÃČES L.: A context-aware method for authentically simulating outdoors shadows for mobile augmented reality. *IEEE Transactions on Visualization and Computer Graphics 24*, 3 (March 2018), 1223–1231. doi:10.1109/TVCG.2017.2676777. 4

[BKC17] BADRINARAYANAN V., KENDALL A., CIPOLLA R.: Segnet: A deep convolutional encoder-decoder architecture for scene segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence PP*, 99 (2017), 1–1. doi:10.1109/TPAMI.2016.2644615. 2, 3

[BUSB15] BELL S., UPCHURCH P., SNAVELY N., BALA K.: Material recognition in the wild with the materials in context database. *Computer Vision and Pattern Recognition (CVPR)* (2015). doi:10.1109/CVPR.2015.7298970. 2, 5

[CPK*17] CHEN L. C., PAPANDREOU G., KOKKINOS I., MURPHY K., YUILLE A. L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence PP*, 99 (2017), 1–1. doi:10.1109/TPAMI.2017.2699184. 2, 3

[CTB*12] CHEVAILLIER P., TRINH T.-H., BARANGE M., DE LOOR P., DEVILLERS F., SOLER J., QUERREC R.: Semantic modeling of virtual environments using mascaret. In *Software Engineering and Architectures for Realtime Interactive Systems (SEARIS), 2012 5th Workshop on* (2012), IEEE, pp. 1–8. doi:10.1109/SEARIS.2012.6231174. 3

[CY00] COUGHLAN J. M., YUILLE A. L.: The manhattan world assumption: Regularities in scene statistics which enable bayesian inference. In *Proceedings of the 13th International Conference on Neural Information Processing Systems* (Cambridge, MA, USA, 2000), NIPS'00, MIT Press, pp. 809–815. URL: http://dl.acm.org/citation.cfm?id=3008751.3008869. 8

[DRMS07] DAVISON A. J., REID I. D., MOLTON N. D., STASSE O.: Monoslam: Real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 6 (June 2007), 1052–1067. doi:10.1109/TPAMI.2007.1049. 2

[DTKPB07] DE TROYER O., KLEINERMANN F., PELLENS B., BILLE W.: Conceptual modeling for virtual reality. In *Tutorials, Posters, Panels and Industrial Contributions at the 26th International Conference on Conceptual Modeling - Volume 83* (Darlinghurst, Australia, Australia, 2007), ER '07, Australian Computer Society, Inc., pp. 3–18. URL: http://dl.acm.org/citation.cfm?id=1386957.1386959. 4

[DZMQ16] DENG H., ZHANG L., MAO X., QU H.: Interactive urban context-aware visualization via multiple disocclusion operators. *IEEE Transactions on Visualization and Computer Graphics 22*, 7 (July 2016), 1862–1874. doi:10.1109/TVCG.2015.2469661. 3

[EFN12] ENDERT A., FIAUX P., NORTH C.: Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2012), CHI '12, ACM, pp. 473–482. URL: http://doi.acm.org/10.1145/2207676.2207741, doi:10.1145/2207676.2207741. 3

[FB81] FISCHLER M. A., BOLLES R. C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM 24*, 6 (June 1981), 381–395. URL: http://doi.acm.org/10.1145/358669.358692, doi:10.1145/358669.358692. 3

[FWL17] FISCHBACH M., WIEBUSCH D., LATOSCHIK M. E.: Semantic entity-component state management techniques to enhance software quality for multimodal vr-systems. *IEEE Transactions on Visualization and Computer Graphics 23*, 4 (April 2017), 1342–1351. doi:10.1109/TVCG.2017.2657098. 3

[GGOEO*] GARCIA-GARCIA A., ORTS-ESCOLANO S., OPREA S., VILLENA-MARTINEZ V., GARCIA-RODRIGUEZ J.: A review on deep learning techniques applied to semantic segmentation. arXiv:1704.06857v1. 2, 3

[GLZR17] GRUBERT J., LANGLOTZ T., ZOLLMANN S., REGENBRECHT H.: Towards pervasive augmented reality: Context-awareness in augmented reality. *IEEE Transactions on Visualization and Computer Graphics 23*, 6 (June 2017), 1706–1724. doi:10.1109/TVCG.2016.2543720. 2, 4

[Goo17] GOOGLE: Arcore - google developers, 2017. URL: https://developers.google.com/ar/. 3

[HFL14] HERMANS A., FLOROS G., LEIBE B.: Dense 3d semantic mapping of indoor scenes from rgb-d images. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on* (2014), IEEE, pp. 2631–2638. doi:10.1109/ICRA.2014.6907236. 6

[Int] URL: http://www.intel.com/content/www/us/en/tech-tips-and-tricks/virtual-reality-vs-augmented-reality.html. 1

[JCSL18] JIAO J., CAO Y., SONG Y., LAU R. W. H.: Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *ECCV* (2018). 3

[JSD*14] JIA Y., SHELHAMER E., DONAHUE J., KARAYEV S., LONG J., GIRSHICK R., GUADARRAMA S., DARRELL T.: Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia* (New York, NY, USA, 2014), MM '14, ACM, pp. 675–678. URL: http://doi.acm.org/10.1145/2647868.2654889, doi:10.1145/2647868.2654889. 5

[KK11] KRÄHENBÜHL P., KOLTUN V.: Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems 24*, Shawe-Taylor J., Zemel R. S., Bartlett P. L., Pereira F., Weinberger K. Q., (Eds.). Curran Associates, Inc., 2011, pp. 109–117. URL: http://papers.nips.cc/paper/4296-efficient-inference-in-fully-connected-crfs-with-ga pdf. 6, 7

[KM07] KLEIN G., MURRAY D.: Parallel tracking and mapping for small ar workspaces. In *Proc. 6th IEEE and ACM Int. Symp. Mixed and Augmented Reality* (Nov. 2007), pp. 225–234. doi:10.1109/ISMAR.2007.4538852. 2

[KS80] KINDERMANN R., SNELL J. L.: *Markov random fields and their applications*, vol. 1. American Mathematical Society, 1980. 6

[KSH12] KRIZHEVSKY A., SUTSKEVER I., HINTON G. E.: Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems* (USA, 2012), NIPS'12, Curran Associates Inc., pp. 1097–1105. URL: http://dl.acm.org/citation.cfm?id=2999134.2999257. 3

[KTY*13] KIM H., TAKAHASHI I., YAMAMOTO H., KAI T., MAEKAWA S., NAEMURA T.: *MARIO: Mid-Air Augmented RealityInteraction with Objects*. Springer International Publishing, Cham, 2013, pp. 560–563. 3

[LRKT09] LADICKÃ¡ L., RUSSELL C., KOHLI P., TORR P. H. S.: Associative hierarchical crfs for object class image segmentation. In *2009 IEEE 12th International Conference on Computer Vision* (Sept 2009), pp. 739–746. doi:10.1109/ICCV.2009.5459248. 6

[LS18] LANDRIEU L., SIMONOVSKY M.: Large-scale point cloud semantic segmentation with superpoint graphs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 4558–4567. 3

[MAMT15] MUR-ARTAL R., MONTIEL J. M. M., TARDÓS J. D.: Orb-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics 31*, 5 (Oct. 2015), 114–1163. doi:10.1109/TRO.2015.2463671. 2

[MHDL17] MCCORMAC J., HANDA A., DAVISON A., LEUTENEGGER S.: Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* (May 2017), pp. 4628–4635. doi:10.1109/ICRA.2017.7989538. 3, 6

[Mic17] MICROSOFT: Ball pit - microsoft store, 2017. URL: https://www.microsoft.com/en-us/store/p/ball-pit/9nblggh4wssp. 3

[NEF12] NORTH C., ENDERT A., FIAUX P.: Semantic interaction for sensemaking: Inferring analytical reasoning for model steering. *IEEE Transactions on Visualization & Computer Graphics 18* (12 2012), 2879–2888. URL: doi.ieeecomputersociety.org/10.1109/TVCG.2012.260, doi:10.1109/TVCG.2012.260. 3

[NFS15] NEWCOMBE R. A., FOX D., SEITZ S. M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE CVPR* (2015), pp. 343–352. doi:10.1109/CVPR.2015.7298631. 2, 3

[NIH*11] NEWCOMBE R. A., IZADI S., HILLIGES O., MOLYNEAUX D., KIM D., DAVISON A. J., KOHI P., SHOTTON J., HODGES S., FITZGIBBON A.: Kinectfusion: Real-time dense surface mapping and tracking. In *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on* (2011), IEEE, pp. 127–136. URL: http://dx.doi.org/10.1109/CVPR.2015.7298631. 2, 3, 5

[NLD11] NEWCOMBE R. A., LOVEGROVE S. J., DAVISON A. J.: Dtam: Dense tracking and mapping in real-time. In *2011 International Conference on Computer Vision* (Nov 2011), pp. 2320–2327. doi:10.1109/ICCV.2011.6126513. 2, 3, 4

[NOBW16] NUERNBERGER B., OFEK E., BENKO H., WILSON A. D.: Snaptoreality: Aligning augmented reality to the real world. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2016), CHI '16, ACM, pp. 1233–1244. URL: http://doi.acm.org/10.1145/2858036.2858250, doi:10.1145/2858036.2858250. 3

[PXZ*15] P.WANG, X.SHEN, Z.LIN, COHEN S., PRICE B., YUILLE A.: Towards unified depth and semantic prediction from a single image. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2015), pp. 2800–2809. doi:10.1109/CVPR.2015.7298897. 3

[QSMG17] QI C. R., SU H., MO K., GUIBAS L. J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017). 3

[RA17] RÜNZ M., AGAPITO L.: Co-fusion: Real-time segmentation, tracking and fusion of multiple objects. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* (May 2017), pp. 4471–4478. doi:10.1109/ICRA.2017.7989518. 3

[RBG07] RAUTEK P., BRUCKNER S., GROLLER E.: Semantic layers for illustrative volume rendering. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (Nov 2007), 1336–1343. doi:10.1109/TVCG.2007.70591. 3

[SLD17] SHELHAMER E., LONG J., DARRELL T.: Fully convolutional networks for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence 39*, 4 (2017), 640–651. URL: http://dx.doi.org/10.1109/TPAMI.2016.2572683. 2, 3, 5, 9

[SMGKD14] SALAS-MORENO R. F., GLOCKEN B., KELLY P. H. J., DAVISON A. J.: Dense planar slam. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (Sept 2014), pp. 157–164. doi:10.1109/ISMAR.2014.6948422. 3

[SWRC09] SHOTTON J., WINN J., ROTHER C., CRIMINISI A.: Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision 81*, 1 (Jan 2009), 2–23. 7

[SZ15] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations* (2015). URL: https://arxiv.org/abs/1409.1556. 5

[TTLN17] TATENO K., TOMBARI F., LAINA I., NAVAB N.: Cnn-slam: Real-time dense monocular slam with learned depth prediction. *CoRR abs/1704.03489* (2017). URL: https://arxiv.org/abs/1704.03489. 3

[WJK*13] WHELAN T., JOHANNSSON H., KAESS M., LEONARD J. J., MCDONALD J.: Robust real-time visual odometry for dense rgb-d mapping. In *2013 IEEE International Conference on Robotics and Automation* (May 2013), pp. 5724–5731. doi:10.1109/ICRA.2013.6631400. 3

[Yan18] YANG W.: Context-aware computer aided inbetweening. *IEEE Transactions on Visualization and Computer Graphics 24*, 2 (Feb 2018), 1049–1062. doi:10.1109/TVCG.2017.2657511. 3

[YEII16] YU L., EFSTATHIOU K., ISENBERG P., ISENBERG T.: Cast: Effective and efficient user interaction for context-aware selection in 3d particle clouds. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (Jan 2016), 886–895. doi:10.1109/TVCG.2015.2467202. 3

[ZCC16] ZHANG E., COHEN M. F., CURLESS B.: Emptying, refurnishing, and relighting indoor spaces. *ACM Trans. Graph. 35*, 6 (Nov. 2016), 174:1–174:14. URL: http://doi.acm.org/10.1145/2980179.2982432, doi:10.1145/2980179.2982432. 8

[ZCX*18] ZHANG Z., CUI Z., XU C., JIE Z., LI X., YANG J.: Joint task-recursive learning for semantic segmentation and depth estimation. In *Computer Vision – ECCV 2018* (Cham, 2018), Ferrari V., Hebert M., Sminchisescu C., Weiss Y., (Eds.), Springer International Publishing, pp. 238–255. 3

[ZJRP*15] ZHENG S., JAYASUMANA S., ROMERA-PAREDES B., VINEET V., SU Z., DU D., HUANG C., TORR P.: Conditional random fields as recurrent neural networks. In *International Conference on Computer Vision (ICCV)* (2015). doi:10.1109/ICCV.2015.179. 2, 3, 5, 6, 9

[ZSS17] ZHAO C., SUN L., STOLKIN R.: A fully end-to-end deep learning approach for real-time simultaneous 3d reconstruction and material recognition. In *2017 18th International Conference on Advanced Robotics (ICAR)* (July 2017), pp. 75–82. doi:10.1109/ICAR.2017.8023499. 3, 6