performance, for example by playing instruments and singing on stage [2].

The second experience was around connecting with others while being apart [L5]. In particular, we showed how our VR2Gather platform could be used to bring people together and share a virtual cake slice. Our setup included a specific capturing system for the cake itself along with the users, which were able to chat and interact in real time while being located in different cities.

The last experience was around remote consultation with doctors [L6]. Meeting remotely in an immersive environment opens new possibilities for healthcare, for example by reducing the amount of time patients spend travelling to the clinic and waiting for their appointment, and by enabling people with mobility impairments to access healthcare advice in real time, while waiting for the healthcare personnel to be dispatched on site. However, patients might not have access to high-end volumetric cameras, or stable connections. Thus, we demonstrated the consultation using a consumer-grade phone to acquire a volumetric representation of the patient, which was transmitted over 5G network.

This work was supported through "PPS-programmatoeslag TKI" Fund of the Dutch Ministry of Economic Affairs and Climate Policy and CLICKNL, the European Commission H2020 program, under the grant agreement 762111, VRTogether [L7] and the European Commission Horizon Europe program, under the grant agreement 101070109, TRANSMIXR [L8].

Links:

[L1] https://www.wired.com/story/gadget-lab-podcast-630/
[L2] https://www.intelrealsense.com/
[L3] https://volucap.com/
[L4] https://kwz.me/hAx
[L5] https://kwz.me/hAz
[L6] https://kwz.me/hAA
[L7] http://vrtogether.eu/
[L8] https://transmixr.eu/

References:

[1] I. Viola, et al., "VR2Gather: A collaborative social VR system for adaptive multi-party real-time communication," IEEE MultiMedia, 2023.
[2] I. Reimat, et al., "Mediascape XR: a cultural heritage experience in social VR," in Proc.of the 30th ACM International Conference on Multimedia, pp. 6955–6957, 2022.

Please contact:

Irene Viola, CWI, The Netherlands
irene@cwi.nl

Bridging Virtual and Physical Worlds through AI

by Fabio Carrara (CNR-ISTI)

Immersive and user-friendly experiences will win the Extended Reality (XR) game in the long run. However, setting up good VR/AR scenarios often requires manual asset authoring, which is realistic just when dealing with a limited number of predefined objects and scenes. The Social and hUman ceNtered (SUN) XR project is investigating low-cost, yet effective, solutions to create links between a physical environment and its corresponding one in the virtual world.

The main gateway to our digital life is still the smartphone, but as technology evolves, we are witnessing a shift towards more immersive and seamless experiences. We spend several hours on the smartphone, utilising its communication, entertainment, and productivity capabilities. However, there are scenarios and tasks where the limitations of the smartphone form factor and interface become clear. XR could fill the gap in scenarios where other interfaces cannot be adopted, offering a more intuitive and immersive way to interact with digital content and information by blending the virtual and physical worlds.

The SUN XR project [L1], funded by the European Commission within the Horizon Europe program and uniting forces of 18 partners from eight countries, aims at exploring the potential of XR in a new era of digital life that better integrate and is symbiotic with the "analogue" life. Specifically, the project focuses on demonstrating how new XR technologies can improve social well-being in diverse contexts, such as healthcare and work environments.

Current XR technology is still in its infancy; setting up a VR or AR experience is a complex and time-consuming task, often requiring digital artists to manually author assets to support interactions with limited, controlled, and non-personalised scenes and objects. One of the objectives of the SUN XR project is to improve the acquisition and understanding of the physical world surrounding us, thus facilitating the transition between the users' physical and virtual environments for general and personalised XR experiences. AI plays a crucial role in achieving this goal; the vast leaps in semantic understanding of the world in open and unconstrained settings can revolutionise how we interact with the digital world and make XR experiences more immersive and user-friendly.

In this context, the CNR-ISTI team is developing novel AI-based methodologies and tools for open-world semantic understanding and multimodal interaction with general environments and objects through sensor streams commonly available in XR devices, i.e. RGB and depth video streams. Two main problems are currently being pursued: understanding dynamic objects (via open-vocabulary object detection) and environment (via scene understanding and 3D reconstruction) semantically. These two tasks could provide the foundation tools for supporting more natural interactions between the user and XR apps.

For the understanding of dynamic objects, open-vocabulary object detection and segmentation models [1] provide a fast

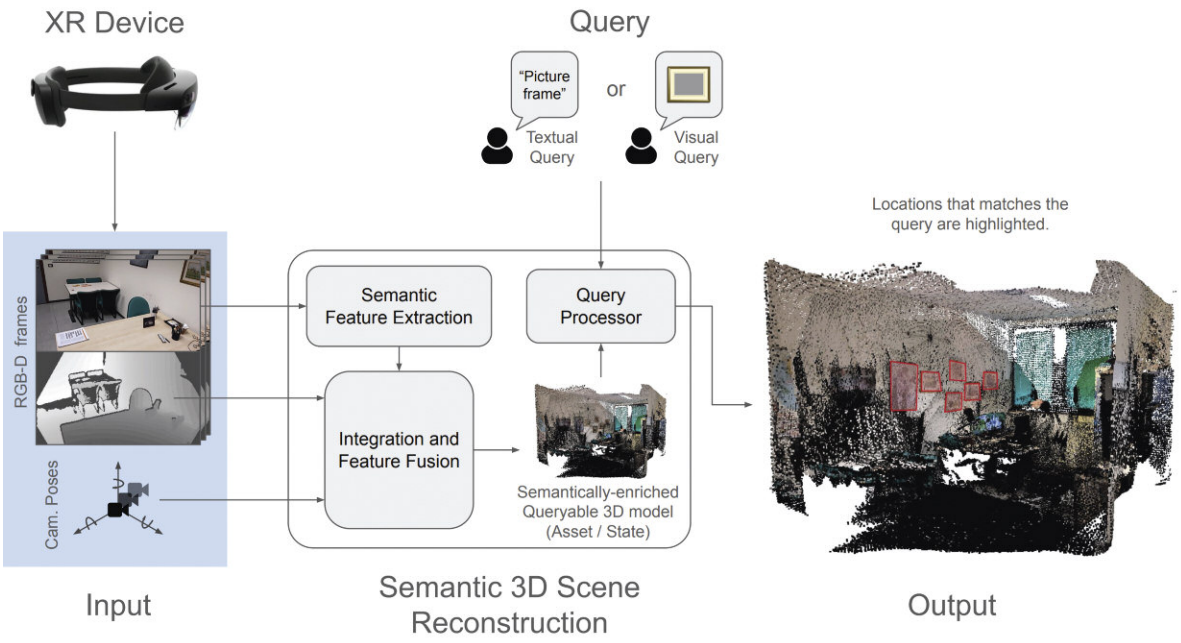

Figure 1: High-level scheme of a semantic-enriched 3D reconstruction pipeline. Semantic understanding is performed on RGB images using state-of-the-art foundation vision-language models and integrated into 3D assets. The user can then interact with its surrounding by textual (e.g. transcribed from speech) and visual (e.g. by gazing, pointing, or external image resources) queries in an open-world setting.

way to locate and recognise objects of interest in the environment without the need for a predefined set of classes. These models exploit the power of general multimodal representations to match regions of the input image with a textual query provided at inference time by the user. This allows for a more natural and flexible interaction with the environment, as the user can ask for objects of interest without the need to know their names in advance. However, open-vocabulary models still have limitations in scenarios where fine-grained object recognition is required, as demonstrated by our current research on evaluating such models in fine-grained description understanding [2, L2]. We deem fine-grained understanding crucial to adapt to complex environments. Our results show that the primary source of errors in fine-grained open-vocabulary detection is the misalignment between the textual query representation and the visual features of the objects, which require more complex matching mechanisms to be solved and are currently under investigation.

For the understanding of the static environment, we are investigating solutions that integrate vision-based multimodal semantic representations (e.g. coming from CLIP-like models [3]) into 3D reconstruction pipelines, providing a semantically-aware 3D representation of the environment that can be queried by free-form natural language queries or even visual queries. Users could ask for objects or scenes of interest via speech and interact with the environment through visual queries, e.g. by gazing or pointing at objects or scenes of interest (see Figure 1). We are currently working on integrating these models into a 3D reconstruction pipeline based on RGB-D data from available AR/VR devices. Current limitations are related to the fusion of multimodal representations from different viewpoints into a coherent 3D representation of the environment, often due to the lack of fine geometric information in the input data, and in making the pipeline efficient and scalable.

We will demonstrate the potential of these AI-based methodologies in two pilots. The first one concerns safety training and innovative cooperation in a manufacturing workplace. The proposed XR systems will provide a more immersive and interactive training experience, allowing workers to interact with the physical environment while receiving real-time feedback and guidance on safety guidelines, such as PPE usage and safe working practices, and a more streamlined overview of the machines and tasks status in the workplace. The second pilot concerns using XR to facilitate interaction with motor- or communication-impaired people. Our semantic 3D reconstruction pipelines and multimodal interaction will allow for a personalised and more familiar immersion in the virtual environment, thus improving the user's communication with their relatives and overall social well-being.

Links:

[L1] https://www.sun-xr-project.eu/

[L2] https://lorebianchi98.github.io/FG-OVD/

References:

[1] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," Advances in Neural Information Processing Systems, vol. 36, 2024.

[2] L. Bianchi et al., "The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding," IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), 2024.

[3] A. Radford et al., "Learning transferable visual models from natural language supervision," Int. Conf. on Machine Learning (ICML), PMLR 139, 2021, pp. 8748-8763.

Please contact:

Fabio Carrara, CNR-ISTI, Italy
fabio.carrara@isti.cnr.it