This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101092889, Topic HORIZON-CL4-2022-HUMAN-01-14

SHARESPACE

Embodied Social Experiences in Hybrid Shared Spaces

Project Reference No	101092889
Deliverable	D3.7. Multi-sensors and Multifocal XR display v1
Workpackage	WP3: Visual-Inertial Body Tracking, Scene Capturing, and XR-Display
Nature	DEM (Demonstrator)
Dissemination Level	PU - Public
Date	6 June 2024
Status	V1.0
Editor(s)	Kristians Slics, Roberts Zabels, So Okumura, Takumi Sato
Involved Institutions	Lightspace, Ricoh
Document Description	This document elucidates development and integration progress to derive an optically see-through XR head-mounted display deployable within "Sharespace" environment.

CONTENTS

LIST OF TABLES

Table 1: List of Abbreviations

Term / Abbreviation	Definition
AR	Augmented Reality
VR	Virtual Reality
XR	eXtended Reality
HMD	Head Mounted Display
NED	Near-to-Eye Display
SOM	System on Module
SOC	System on Chip
HW	Hardware
SW	Software
FW	Firmware
VAC	Vergence-Accommodation Conflict
IMU	Inertial Measurement Unit (Gyroscope, Accelerometer, Magnetometer)
SLAM	Simultaneous Localization and Mapping
VIO	Visual-Inertial Odometry
FPGA	Field Programmable Gate Array (A type of integrated circuit)
FOV	Field-of-View

1 INTRODUCTION

This document displays the substance of project's SHARESPACE deliverable 3.7 – *Multi-sensors and Multifocal XR display v1*. The corresponding deliverable is related to development of optically see-through extended reality (XR) head-mounted display (HMD) platform for deployment under SHARESPACE use-case scenarios. In this case the HMD can be understood as augmented reality (AR) near-to-eye display (NED), which is utilized to overlay digital 3D content onto the real-world – thus augmenting the reality by otherwise unavailable information. Such way of displaying 3D information is beneficial as it doesn't isolate the viewer from the real-world – the viewer is still in the reality but at the same time can sufficiently immersively perceive contextual digital content – for example perceive effects of in-person socialization based on observed characters.

In the first segment of the project, version 1 of the headset system is achieved to demonstrate foundational hardware and related software/firmware functionality. Work on the HW is carried out by two project partners – Lightspace, which develop the screen technology for human-centric visualization and Ricoh, which develop proprietary eye (gaze) tracking subsystem to be integrated within the AR HMD as the part of the work package 3.

This report covers taken steps towards achieving necessary functionality as well as briefly illustrates further steps towards *version 2* of the HMD platform. A short video illustrating essence of the deliverable is available on the YouTube platform: https://www.youtube.com/watch?v=BhnsUvz-Hro

2 BACKGROUND

2.1 STEREOSCOPIC DISPLAYS AND POTENTIAL PROBLEMS

The deliverable 3.7 is a part of the hardware component of the SHARESPACE project, with which the corresponding use-case scenarios are visualized. There are certain factors to be taken into account to ensure safe human-centric experiences. Stereoscopic displays are known way to convey a 3D image but usually – as in the case of most readily available systems – weather of VR or AR type, the stereoscopic displays utilize a single focus plane. Typically, this focus plane is strategically located at around 2-2.5 meters which is optimized for certain type of VR games. Nonetheless, people with functional stereoscopic vision, when perceiving 3D world rely on various depth cues which are working in consonance – these include monocular focus cues (conveyed through eye accommodation) as well as eye vergence cues – related to the binocular disparity. Thus, to observe a single sharp image, we have to focus and verge our eyes at target point in the real world. This is the natural expected state by our brains. In a conventional stereoscopic display this condition is broken due to availability of single focus plane. That is – in order to observe a sharp image – eyes have to be focused (accommodated) at a fixed distance (at which the screen plane is located), while eye vergence is engaged to ensure single image. The binocular disparity is used as the strongest 3D depth cue which drives eye vergence. While this fools the brain and we perceive the image to have 3D properties, the mismatch in the vergence distance and the accommodation distance induces vergence-accommodation conflict (VAC). The expected condition for single sharp image is broken. The VAC Can manifest as eyestrain, fatigue, nausea, blurred vision etc (David M. Hoffman, 2008). In short – it is very unfavorable condition, which should be avoided at all costs – especially in the context of what a Sharespace platform is intended to offer. An example showing principles of stereoscopic image is shown in Figure 1.

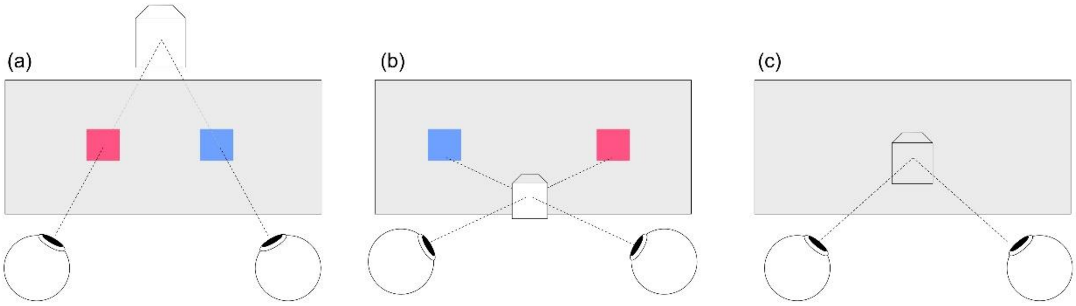

Figure 1. A mode in which a stereoscopic display can convey a 3D image is by showing two distinctly different images – each to the corresponding eye, where each of the images shows a scene or a 3D object from a slightly different perspective. Thus, if there is available only one image focus plane – various conditions in terms of disparity between eye accommodation and vergence cues can occur. If the object is displayed beyond the focus plane – an uncrossed displarity or VAC is observed [case (a)]. If the object is displayed in front of the focus plane – a crossed disparity or VAC is observed [case (b)]. And if the 3D object is displayed exactly on the focus plane – no disparity is observed [case (c)]. The case [c] is ab expected condition when observing real-world but a rare occasion in a stereoscopic 3D image space.

2.2 MULTIFOCAL OPTICAL ARCHITECTURE

A solution undertaken within the scope of Sharespace project is a "multi-focal NED optical architecture" (Zabels, u.c., 2019). The concept foresees proprietary core image-forming engine developed by Lightspace. In essence – instead of a single focus plane, a multi-focal optical architecture presents the 3D content over several focus planes – thus enabling arbitrary refocusing (accommodation) of eyes within a 3D scene. By introducing several focus planes – it is possible to match the eye vergence and accommodation cues such that the disparity is below the threshold of sensitivity and no negative effects are occurring. That is – with strategic placement of image focus planes it becomes possible to keep the VAC below certain value – 0.3-0.4 Diopters, which has been shown as a safe threshold (Takashi Shibata, 2011). An example of such focus plane layout is shown in Figure 2.

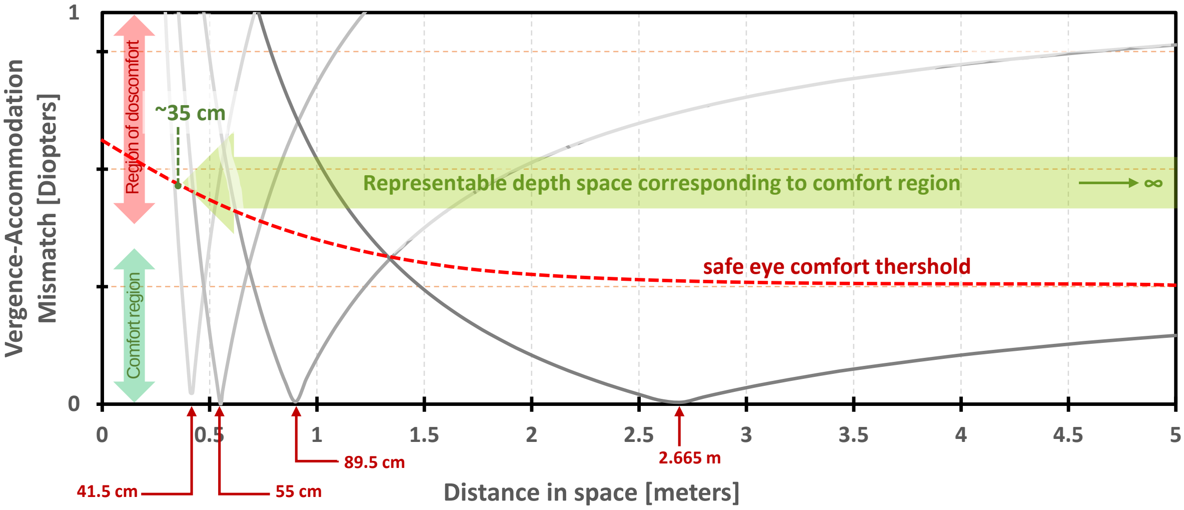

Figure 2. An example of Lightspace's multi-focal NED optical architecture based on 4 focus planes. With such layout – placement of focus planes at 0.41m, 0.55m, 0.9m and 2.67m full eye comfort is guaranteed from around 35 cm to infinity. The image shows vergence-accommodation mismatch for said focus planes as a function of distance. The crossing points indicate the distance at which the 3D content is split between corresponding focus planes-thus ensuring the best match of vergence and accommodation depth cues.

In practice Lightspace NED or HMD optical architecture relies on optical diffuser technology and image back projection, see Figure 3. A high image refresh-rate micro projection unit is driven at 240 Hz or more (corresponding to 4-focus plane implementation) and projects portions – focus planes – at a multi-layer screen element formed by switching diffuser elements. In this context – the switching diffuser elements is a secondary image source ensuring distinct focal distances (As shown in Figure 2) when combined with a magnifying eyepiece. The system operates time-sequentially and scans the 3D volume – but due to high repetition rate, which is above typical flicker fusion frequency, a viewer perceives the image as having true focal depths. On one hand this approach solves vergence-accommodation conflict, but of equal importance is the absence of focal rivalry in an optically see-through AR mode. That is – the virtual content can be matched to real-world adding to immersivity and excluding spatial/focal confusion, which can occur with single focus-plane optically see-through devices.

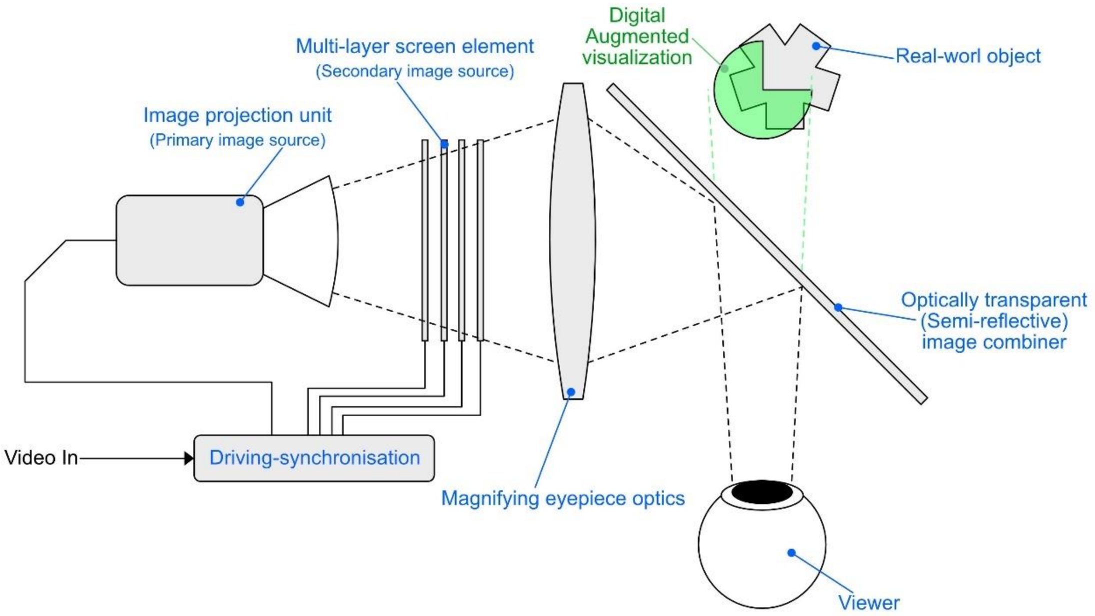

Figure 3. Simplified block schematic of a multi-focal NED implementation in an AR-modality. Shown is only a single channel (single eye). The primary image source is a micro projection unit operating at a high frequency (240 Hz), while the secondary image source is liquid crystal switching elements – driven time-sequentially, thus scanning the whole 3D volume of a representable 3D image. Due to high repetition rate, from the perspective of a viewer – all focal depths are available simultaneously – providing full freedom to refocus the eye.

2.3 KEY SENSOR COMPONENTS OF XR HMD SYSTEM

To ensure a meaningful XR experience, the HMD platform has to encompass not just the display element but various additional sensor modalities, which form the basis for contextually accurate and organic user experience. The main components are:

- Display
- Pose tracking system
- Means of audio recording
- Means of audio playback
- Means of interaction (hand tracking, gesture recognition, voice commands, remote control)
- Eye or gaze tracking

2.3.1 Pose tracking system

Pose tracking is the key to contextually accurately position the virtual content in respect to real world. These are means how to track the viewer in respect to the surroundings (real world). In actuality tracked is the HMD, assuming the viewer is semi-rigidly positioned in respect to the HMD’s display. Whereas accuracy of image in respect to the viewer is achieved either by preliminary display calibration to a particular user and/or involvement of other means – namely gaze tracking.

There are two main groups of pose tracking approaches – the Inside-out and the Outside-In. A more versatile option is the Inside-out variant – as it operates just like humans – the whole set of sensors is bult into the HMD and it “observes” the surrounding to calculate the origin point and relative/absolute movement with 6 degrees of freedom. Nonetheless, accuracy of this method is strongly dependent on SW/FW solutions and available computational resources. The pose has to be calculated in real-time and due to inevitable signal delays (latency) strategies of pose-prediction have to be also employed to forecast actual pose of the headset ahead of the time. Typical computational methods utilized for this approach are simultaneous localization and mapping (SLAM) – most commonly relying on wide-angle stereoscopic cameras and IMU sensor.

The alternative Outside-in approach relies on external (fixed) tracking units – most widely also camera systems, which are calibrated and are looking for a position of a HMD within the 3D space. Nonetheless, in the context of Sharespace, this is a less favorable approach as the tracking volume is limited, the setup is complex and oftentimes expensive.

2.3.2 Audio input/output

Bidirectional audio interface is a self-explanatory need to ensure verbal communication and perception of other audial information within Share space scenarios. A single or set of high sensitivity microphones as well simple high-def speakers ensures this common and naturally expected feature.

2.3.3 Means of interaction

For more meaningful experience a mode of interaction with the digital component of the world has to be also ensured. A straight forward approach can be a HW-based manipulator – a remote, a keyboard. Nonetheless, such approach can disrupt the organic feel of the experience and more natural means of interaction have to bee investigated. Also, quite commonly a hand tracking is employed together with gesture recognition – For this dedicated, also wide-angle, cameras can be used.

2.3.4 Eye/gaze tracking

In the context of Sharespace scenarios and animation of digital avatars and important role is devoted to realistic eye animations which require real-time knowledge about the viewer's gaze. This is an important component which can be implemented in various ways – most widespread of which is dedicated tiny infrared cameras and corresponding infrared LED illumination. Nonetheless, such approach can violate privacy and security, as captures video datasets. Instead within the Sharepsace project a partner Ricoh provides a solution based on IR laser beam illumination and registration of gaze direction without direct capture of identifiable video material – by means of capturing laser-beam deviation on a 2D sensor.

3 APPROACH

3.1 FOUNDATIONAL HMD

The foundation for the V1 XR HMD is based on Lightspace's IG1050 series headset, which has been developed for medical use-cases. It is a 4-focus plane NED in a full-metal construction that is tethered to a host PC, Figure 4.

For the Sharespace V1 XR HMD, the optical train has been maintained – it is based on so-called "bird-bath" optical combiner, which is a reflective-type optics with well controlled color aberrations and high image quality due to on-axis design, Figure 5. Nonetheless, overall design changes concerned weight and power consumption reduction. The body material was changed from full metal to polycarbonate. Alongside this amendment, main board design has been changed to utilize lower-power components – including the main FPGA. Overall, this approach enabled saving of more than 30% in weight and reduce the power consumption by 25%.

In the course of further integration actions changes were introduced to mechanical components to accommodate eye-tracking modules.

Figure 4. Multi-focal (4-focus planes) IG-1050 AR HMD by Lightspace.

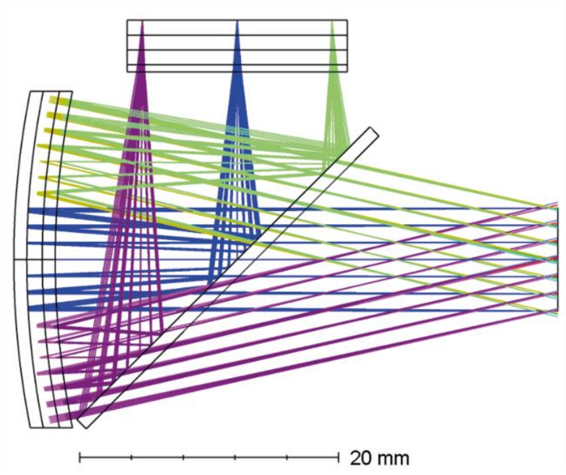

Figure 5. "Bird-bath" optical image combiner-magnifyiong eyepiece. A cross section from Zemax Optic Studio. The image source is the multi-layer screen element (on top). In such configuration the real-world transmission typically is around 25%.

A list of key parameters for the V1 HMD system is provided in Table 2.

Table 2. Key technical parameters of V1 HMD system

Specification	**Sharespace XR HMD V1**
Optical architecture	Solid-state multi-focal
Image source	Pico projection unit
Number of focal planes	4
Image resolution	FWVGA
Image refresh rate	60 Hz (volumetric)
Image FOV	42-deg (H), 24-deg (V), 48.5-deg (D)
Image brightness	420-550 nits per depth plane
Eye-box	14x8 mm (without image cropping)
Eye relief	default value 18 mm
See-through transparency	<25%
Connection to PC	tethered - USB-C through port-combiner (power supply)
OS support	Windows 10
Weight	1058g (full weight, incl. cable), 430g (front visor part)
Power consumption	<15W (typ.)
Optional integrated features	In-house calibrated IMU, stereo cameras, microphone, loudspeakers,
Pose tracking	Optical outside-in supported, Inertial visual odometry, v-SLAM (experimental)

3.2 EYE TRACKING SUB-SYSTEM

The missing component for the V1 HMD system is the eye tracking – a crucial part for accurate animations as well as supplementary means to improve image rendering and possibly provide a layer for the interaction with the digital content.

For this purpose, a proprietary eye tracking solution developed by Project partner – Ricoh was chosen to be integrated within a HMD platform, Figure 6.

A photograph of version 1 sensor arrangement as to be fitted on a headset is shown in Figure 7.

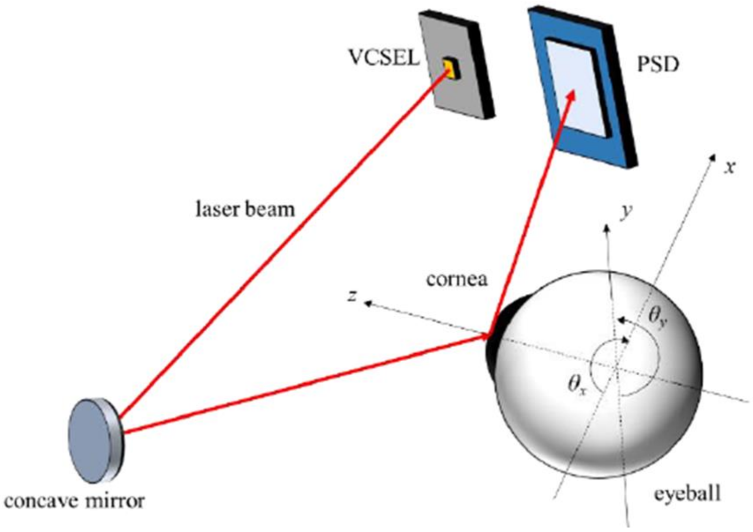

Figure 6. Basic schematic of a single eye-tracking sensor arrangement as proposed by Ricoh. The illumination source is a vertical cavity surface emitting laser diode in the infrared (not visible to a human eye) spectrum. The emitted beam is directed towards an eye by reflecting off of a mirror – which also provides a freedom of adjustments. After the second reflection off of cornea the laser beam is directed towards a PSD sensor – a 2D photosensitive array registering the position of laser beam deviation – which corresponds to angular eye movements.

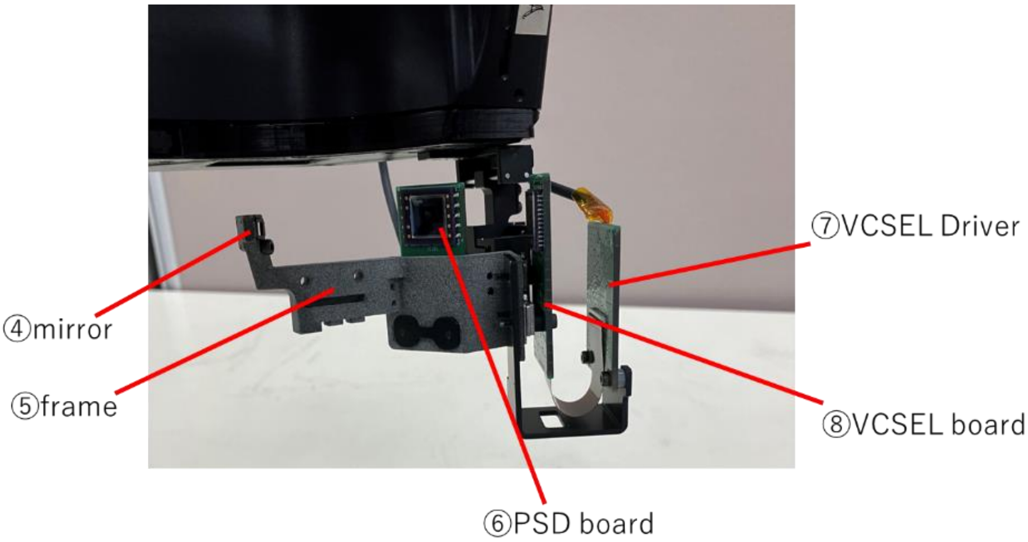

Figure 7. A photograph of Ricoh's VCSEL-based eye tracking prototype unit – showing key components and their relative arrangement.

Core technical parameters of Ricoh’s eye tracking system are provided in Table 3.

Table 3. Core technical parameters for eye-tracking subsystem by Ricoh

Specification	**1st prototype (May 2024)** *per eye
Eye tracking method	Sensor-based eye tracker without camera
Output data	Time stamp, gaze angle
Sampling rate	1kHz
Field of view	±13 deg (Horizontal) *preliminay ±13 deg (Vertical) *preliminay
Accuracy	< 3deg *preliminay
Resolution	0.2 deg
Interface	USB Type C USB to SPI
Supported OS	Windows10
Power consumption	0.63 W
Device size	Relay board size: 113x40x51mm Mirror size: 4.3x4.3x2mm Frame size: 94x66x72mm
Wavelength	780 nm
Weight	49g（Ricoh parts(only eye tracking device), without USB cable）
Laser class	Class1

The eye-tracking module is designed to fully accommodate comfortable range of natural rotational movements of the eye. Furthermore, the high update rate of 1 kHz enables registration of microsaccadic eye movements – change in which can provide information about the state of the user – for example can be an indication of tiredness or lost alertness.

3.3 OVERALL V1 HARDWARE ARCHITECTURE

In the context of V1 HMD – the main board of the HMD integrates with stereoscopic cameras with a global shutter – offering versatility in terms of use case – but notably to be applicable as a data source for pose tracking algorithms. Likewise essential sensors – in-house calibrated IMU sensors, microphones and stereo speakers are provided. The eye tracking modules from Ricoh in the V1 are integrated on a PC level – that is the eye tracker and the HMD have separate driving circuitries and the integration is on a mechanical side. The HMD is connected to a host computer via USB-C cable, utilizing power delivery, DisplayPort for video data transfer and USB-3 for backflow of sensor data. Similarly, the eye tracker driver is also connected to a same host computer via USB. The corresponding layout is shown in Figure 8.

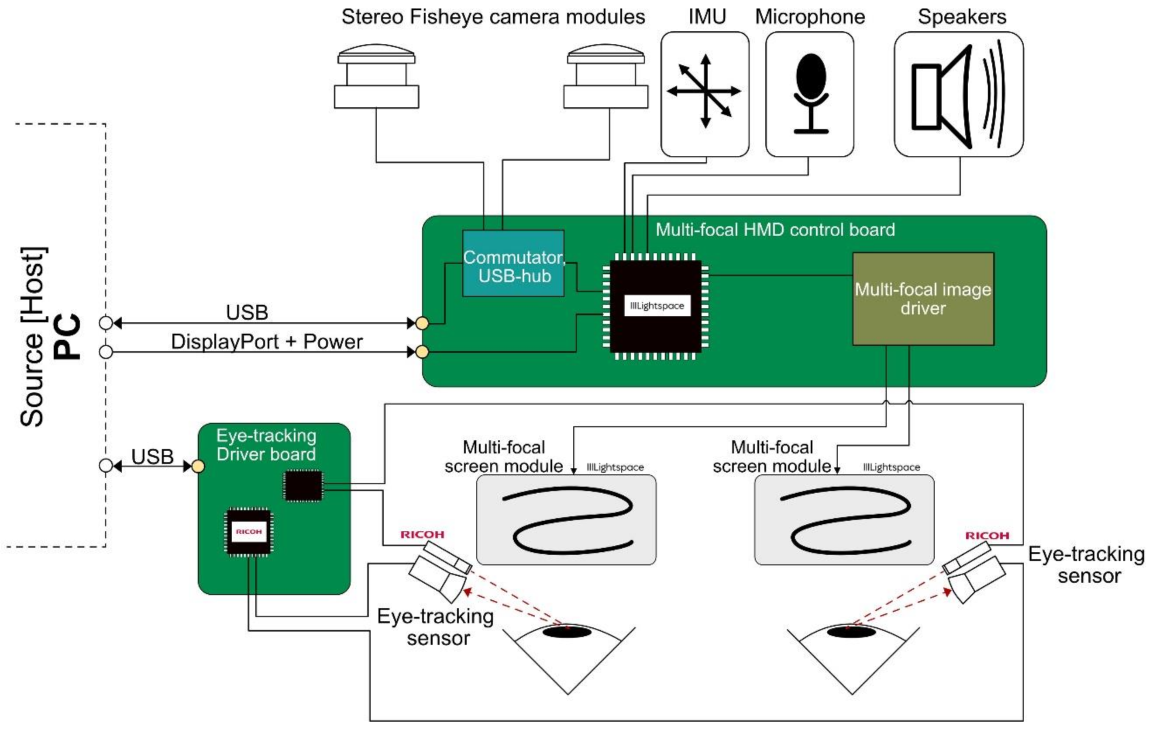

Figure 8. Simplified block-schematic of the XR HMD V1 electronic integration.

3.4 POSE AND HAND TRACKING HARDWARE

To fulfill the need for outside-in pose tracking which can double as a hand tracking camera, built in camera modules were quipped with high-resolution (meaning optical resolution) fisheye lenses, Figure 9.

With a diagonal field-of view of more than 180 degrees – on a given image sensor, it can be possible to avoid inclusion of separate set of cameras dedicated specifically for hand tracking.

In terms of outside-in pose tracking two main approaches were investigated – visual-inertial odometry and visual SLAM algorithms. The difference between these methods is that SLAM algorithms form a global map and can re-localize after lengthier periods when recognizing previously recorded features – but computationally are more demanding. In the course of validation several known approaches were investigated – Bassalt (V. Usenko, 2020), Orbslam3 (C. Campos, 2021) as well as Kimera-VIO (A. Rosinol, 2020).

Figure 9. Fisheye lenses installed on a XR HMD V1 as used for the pose tracking and hand tracking tests.

3.5 OPENXR PLUGIN

The multi-focal HMD by Lightspace has unique way of rendering the image. As of beginning the project – support was ensured only for Unity-based projects. Nonetheless, Sharespace incorporates Unreal engine projects – thus prohibiting the use of multi-focal HMD within Sharespace scenarios.

To achieve full functionality and possibility to use HMD with Unreal projects – OpenXR plugin is being developed. With this standard methods defined in the OpenXR standard will be available to integrate the multi-focal HMD within the software architecture of Sharespace.

4 RESULTS

HMD – base with improved ergonomics, reduced weight, multi-sensor capabilities (Figure 10).

Figure 10. Appearance of a multi-focal HMD as used for the V1 demonstration.

Mechanical integration with eye tracking modules (Figure 11).

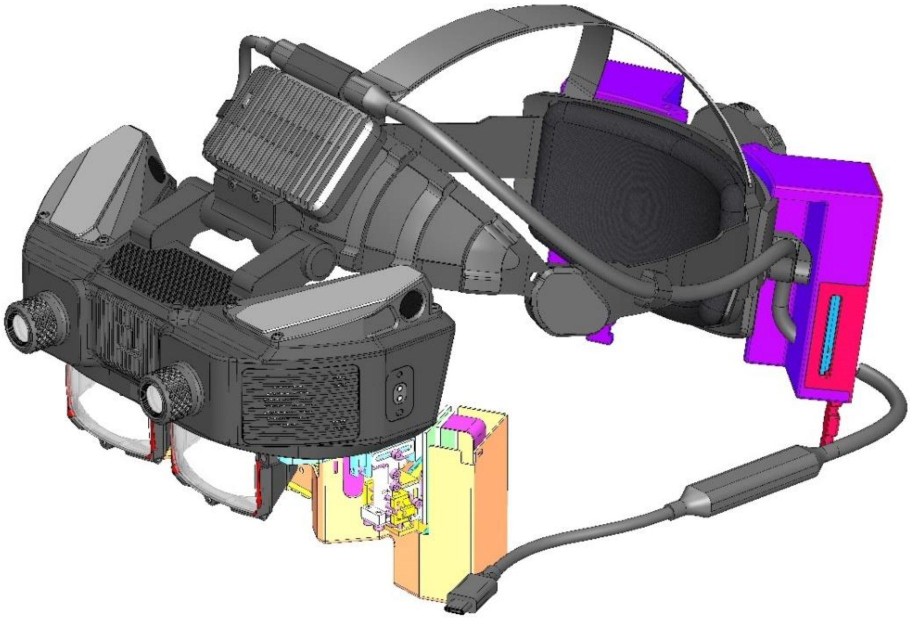

Figure 11. A CAD model of mechanically integrated HMD and the eye-tracking subsystem.

The mechanical integration and development of corresponding components occurred separately in different teams (Ricoh [Japan] and Lightspace [Latvia]). The action has been concluded from the development standpoint – nonetheless final integration step and validation in a laboratory by merging functional eye tracking modules and a functional headset is scheduled for late June 2024, when in dedicated sessions this step will be completed at Lightspace R&D laboratories – working together on-site with engineers from Ricoh.

Within the given work package, integration actions have been carried out on mock-up HMD devices (Figure 12) in Japan.

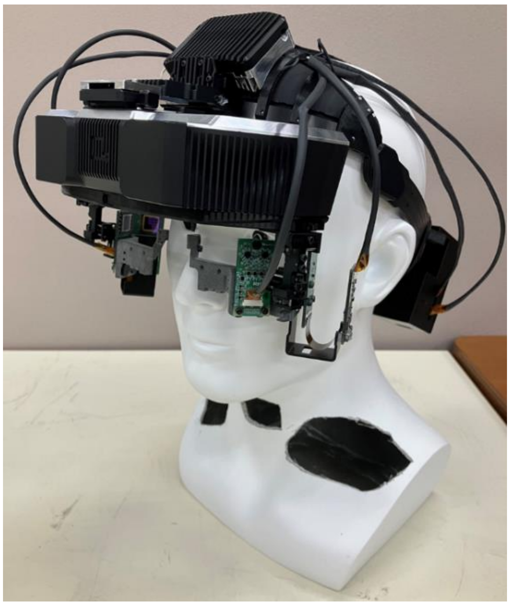

Figure 12. A mock-up HMD system with installed eye tracking modules.

Inside-out tracking

Validated were several wider-known SLAM / VIO open-source models (Basalt, Orbslam3 and Kimera with and without closed loop detection) in different modes. What we present is the translation and rotational error from ground truth at different conditions. Compared were standard lenses, fisheye lenses – purely on visual models and merging with IMU datasets.

In Figure 14 shown is an example of Basalt performance with standard FOV lens and fisheye lenses – furthermore the role of IMU is also assessed. As one would expect – the worst performance is observed for pure visual method using standard FOV lenses. Fishele lenses provide substantially better results while addition of IMU datasets improve on the accuracy even more. On average – in the best case scenario it was possible to achieve less than 2 cm translational error and on average around 0.6 degree rotational error. Nonetheless, the unoptimized version of the Basalt is quite demanding on Central Processing Unit's resources – loading it up to 80%. Some effort was undertaken to optimize the model which reduced the CPU utilization below 60% without loosing performance.

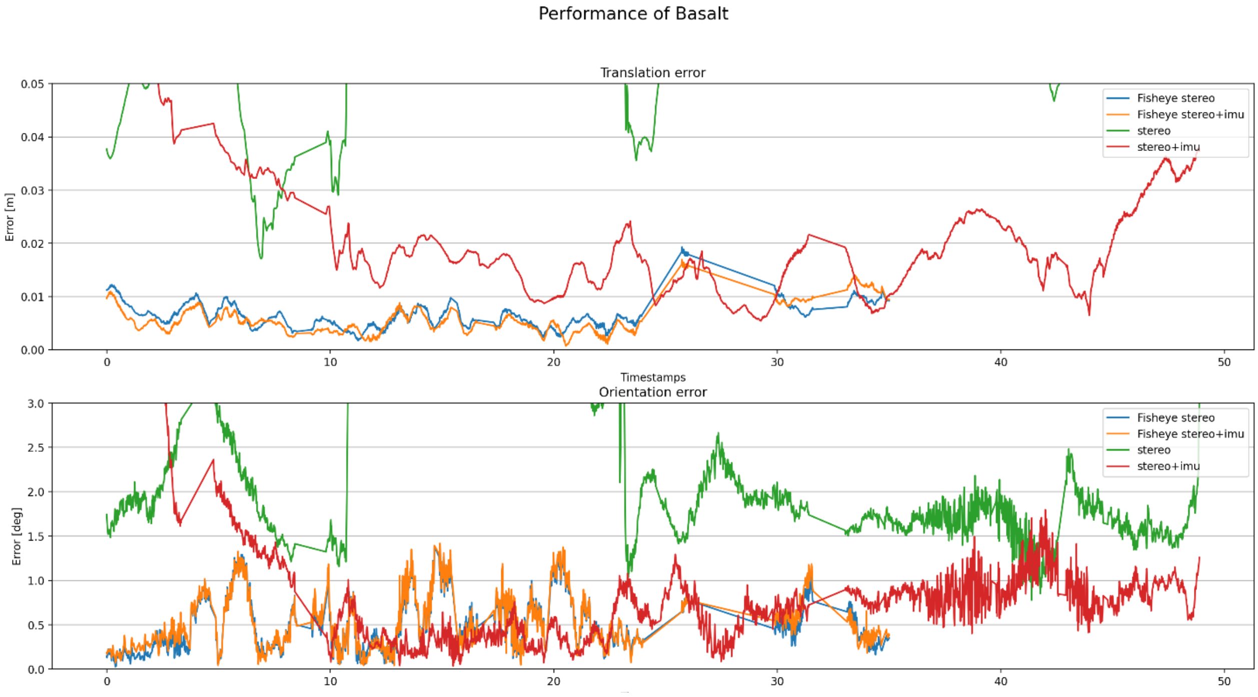

Figure 14. Performance of Basalt model. Compared are video capture with standard lenses (60-degree FOV) and fisheye lenses (198-degree FOV). In addition, evaluated is role of additional IMU data merging. Data gathered with limited movement – using robotic arm as a ground-truth.

A different performance has been observed in case of larger movement registrations – closer to real use-case scenarios, for which the ground truth is taken from a calibrated outside-in tracking system (*Optitrack*). Here, in Figure 13, comparison between 3 verified models can be seen. The translational error can surpass multiple centimeters – especially on abrupt pose changes, while rotational error component remains quite stable – around 0.6-0.8 degrees on average. The difference between

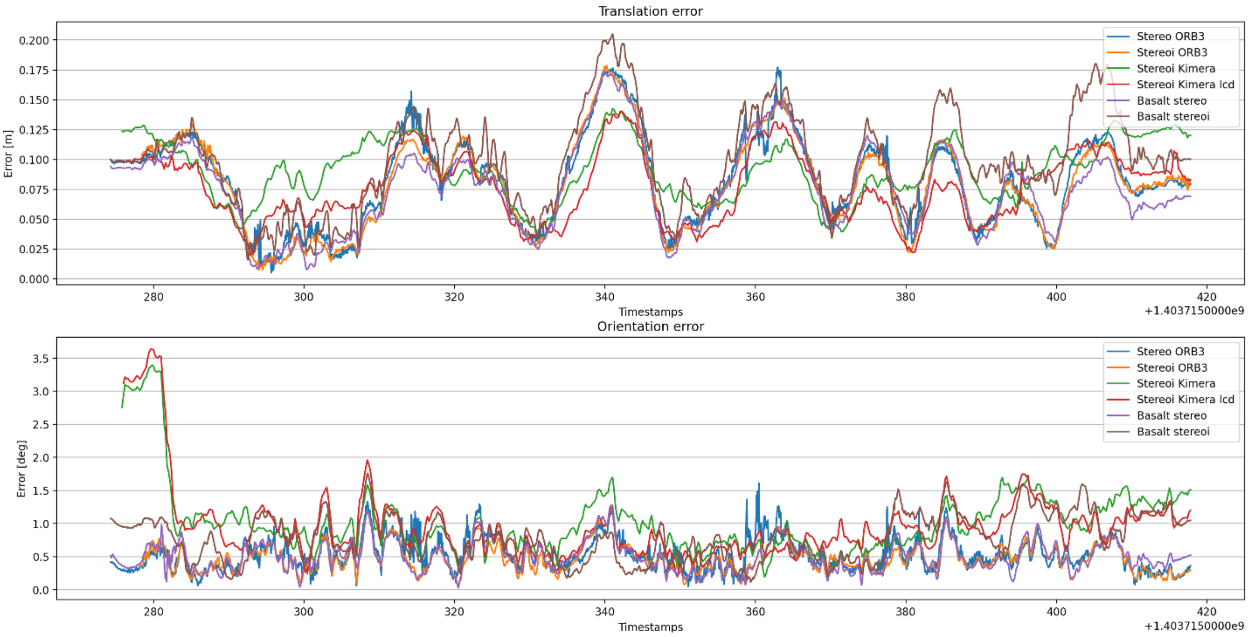

Figure 13. Comparison of Basalt, Orbslam3 and Kimera VIO methods in case of realistic large amplitude movements – including abrupt pose changes. Compared are purely visual methods and methods with addition of IMU data. For VIO, as the name implies IMU data is always part of the loop, the second condition is with a loop close detection.

different methods is changing such that it can be said – they perform comparably. Nonetheless, slightly better performance might be attributed to ORB3 model.

Hand-tracking

Verified was the possibility to use the same camera modules with the installed fisheye lenses for hand tracking feature. As the FOV for this camera is large – it fully overlaps with the FOV of the XR display and extends well beyond – within the region for comfortable hand movements, Figure 15, Figure 16.

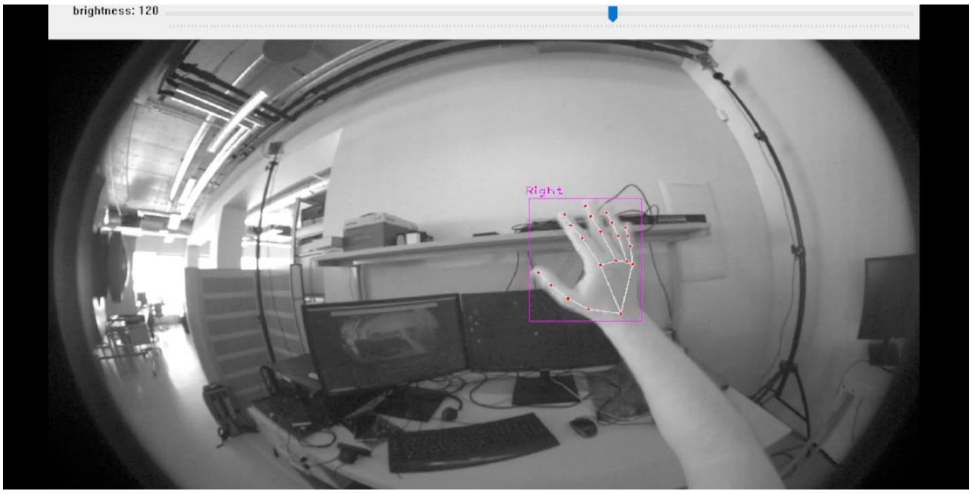

Figure 15. Actual view of a fisheye camera lens as used for hand tracking routines.

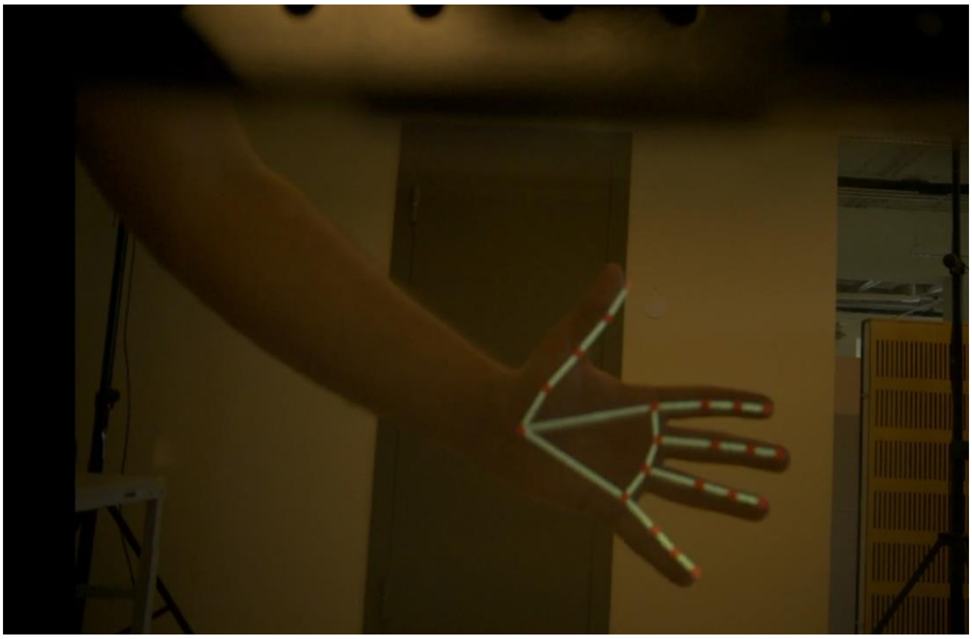

Figure 16. "Through-the-lens" capture of hand-tracking overlay as it can be seen through XR HMD.

5 FURTHER STEPS

The achieved HMD system demonstrates fundamental operation. Nonetheless, there are several directions into which further development is foreseen to take place. In fact, already during development of V1, some of the concepts have been investigated to a certain extent and partly implemented.

OpenXR plugin

Currently the method implementation has been finalized and the wok concerns debugging with subsequent validation – it is foreseen to initially finalize this step (providing fundamental operation) in mid-summer 2024.

Electronic architecture

It is foreseen to take the electronic architecture towards a wireless solution. For this a mini computer or a system on chip would be communicably coupled to a HMD display section. Also in this architecture it would become possible to connect cameras directly to SOM via MIPI-CSI interface – thus widening choice of camera modules. Also – additional cameras could be provided – for example for dedicated hand tracking. A variant of such architecture is shown in Figure 17. Basic steps and initial validation of this concept has already taken place.

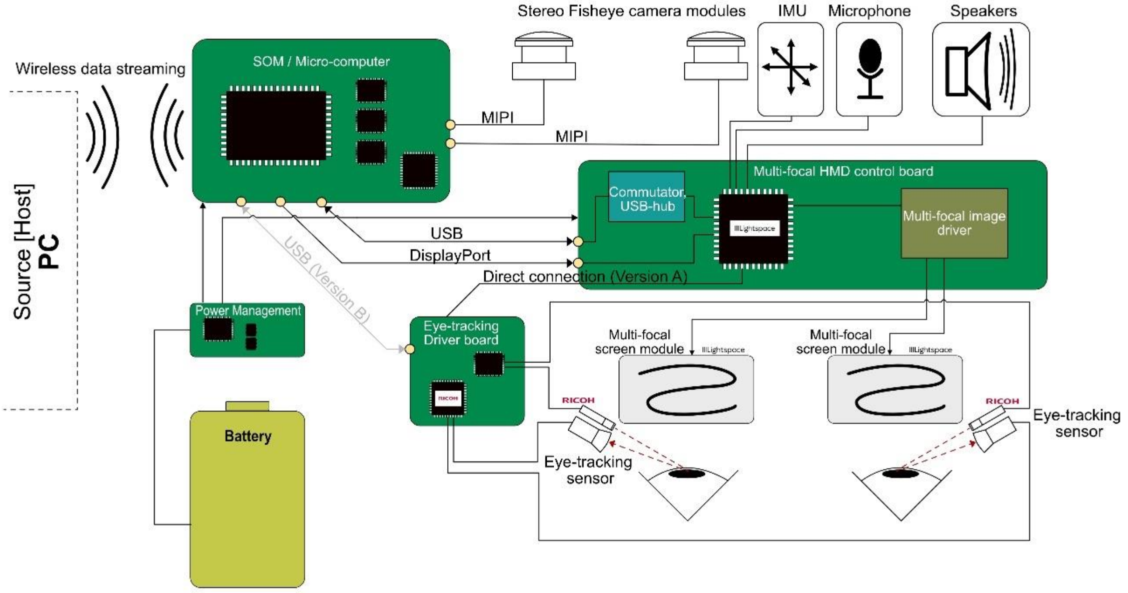

Figure 17. Option of electronic architecture as it could be provided for V2. A wireless solution, where data are streamed wirelessly from a host PC.

Optical architecture

While “bird-bath” image combiner (Figure 5) provides good image quality, it has some drawbacks – namely the see-through transparency is around 25%, optical efficiency for the image source is low (around 10%) and there is an eye glow or front projection – meaning user eyes cannot be seen and other can see the content as well. Thus, alternative optical architecture is being investigated. Instead of a “bird-bath” combiner it is foreseen to choose a flat angled beam-splitter, which also would provide better peripheral vision and an aspheric refractive eyepiece, Figure 18.

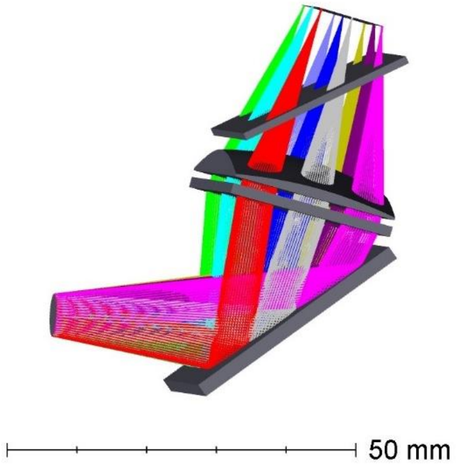

Figure 18. Zemax model of alternative image forming optical pipeline based on aspheric eyepiece and angled flat beam splitter as an optical image combiner.

Pose tracking

During testing phase, it has been established that there is a potential of improving tacking performance. This is related to a type of IMU sensors used. By changing an IMU sensor, it is possible to obtain more stable IMU datasets, which directly and substantially improve stability of pose tracking – overall accuracy as well as jitter. Several IMU sensors with intended manufacturer’s use case in AR/VR are being evaluated.

6 REFERENCES

A. Rosinol, M. A. (2020). Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping. *IEEE Intl. Conf. on Robotics and Automation (ICRA)* (p. 1689). Paris: IEEE. doi:10.1109/ICRA40945.2020.9196885

C. Campos, R. E. (2021). ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. *IEEE Transactions on Robotics, 37*(6), 1874. doi:10.1109/TRO.2021.3075644

David M. Hoffman, A. R. (2008). Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision, 8*(3), 33. doi:10.1167/8.3.33.

Takashi Shibata, J. K. (2011). The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of Vision, 11*(8), 11. doi:10.1167/11.8.11.

V. Usenko, N. D. (2020). Visual-Inertial Mapping With Non-Linear Factor Recovery. *IEEE Robotics and Automation Letters, 5*(2), 422. doi:10.1109/LRA.2019.2961227

Zabels, R., Osmanis, K., Narels, M., Gertners, U., Ozols, A., Rūtenbergs, K., & Osmanis, I. (2019). Option of electronic architecture as it could be provided for V2. A wireless solution, where data are streamed wirelessly from a host PC. *Applied Sciences, 9*(15), 3147. doi:10.3390/app9153147