

UNIVERSITY OF SOUTHAMPTON
Faculty of Engineering and Physical Sciences
School of Electronics and Computer Science

A project report submitted for the award of
MEng Electrical and Electronics Engineering

Supervisor: Dr. Tom Blount

Examiner:

**Open-Source Stereo Video Camera
System and Software
Implementation for Virtual Reality
Lifelogging and Content Creation**

by **Muhammad Hazimi Bin Yusri**

December 7, 2023

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

A project report submitted for the award of MEng Electrical and Electronics Engineering

by Muhammad Hazimi Bin Yusri

REDO SOON In the realm of virtual reality (VR) and lifelogging, this project endeavors to overcome barriers of exclusivity and cost by developing an open-source, low-cost, and modular stereo video camera system. Designed to be modular with the first design to stay on top of a cap, this system integrates lightweight cameras and a microphone with the Raspberry Pi Pico microcontroller. It offers efficient stereoscopic (3D) video capture and immersive surround sound recording. Complementing the hardware, the project entails the development of lifelogging VR software using the Godot game engine. This includes a side-by-side (SBS) video player and intelligent metadata auto-tagging through scene and object detection. The primary objective is to democratize VR content creation, making it accessible to a broad audience, from VR enthusiasts to content creators, encouraging innovation in VR and lifelogging. Challenges, such as technical complexities and power management, are addressed through rigorous prototyping and optimization, ensuring project success and fostering inclusivity, innovation, and the advancement of VR content creation technology in the field of lifelogging.

Contents

1	Introduction	1
1.1	Problem Statement	1
1.2	Goals	2
2	Background and Report of Literature Search	3
2.1	Lifelogging: A Methodical Approach to Memory Preservation	3
2.2	The Significance of Virtual Reality (VR)	3
2.3	Windowed Viewing: Balancing Realism and Feasibility	4
2.4	VR Software for Content Browsing: Maximizing Interaction and Accessibility	4
2.5	Motion Sickness in VR	5
2.6	Examples of Existing Technologies for Lifelogging	5
3	Report on Technical Progress	7
3.1	MoSCoW Requirement	7
3.1.1	Hardware Development	7
3.1.2	Software Preprocessing of Video	8
3.1.3	VR Software Application	8
3.2	Hardware System	8
3.2.1	Cost Analysis	8
3.2.2	Microcontroller and Ecosystem	9
3.2.3	Constraints	9
3.2.4	Onboard Embedded Software Algorithm	9
3.3	Video Pre-Processing Software	11
3.3.1	Motivation	11
3.3.2	FFMPEG for Mono to Stereo Stitching	12
3.3.3	Object and Scene Detection	13
3.3.4	Custom Metadata Tagging	13
3.4	Video Player VR App	13
3.4.1	Game Engine – Godot 4.1	13
3.4.2	Shaders	14
3.4.3	User Interactions (UI)	14
3.4.4	File Browsing	14
	Bibliography	17

Chapter 1

Introduction

1.1 Problem Statement

The recent proliferation of Virtual Reality (VR) Head Mounted Display (HMD) technologies has ushered in a new era of creativity, offering users immersive experiences. However, the formidable barriers of exclusivity and high costs associated with current solutions for recording stereoscopic/3D/spatial video constrain the full potential of these technologies. This issue assumes critical significance for the broader adoption of VR, as low user retention on the platform is often attributed to the scarcity of exclusive content. The redundancy of viewing conventional 2D content through VR headsets, given the superior displays and audio capabilities of other mainstream devices such as TVs, phones, and tablets, underscores the urgent need for a solution. Empowering the general consumer to effortlessly create their own stereoscopic video content not only enhances the appeal and novelty of VR but also addresses the demand for exclusive and engaging material.

Traditional approaches to VR content creation often prioritize factors like Field of View (FOV) and 3 Degrees of Freedom (3DoF), resulting in 180-degree or 360-degree videos. However, these videos often lack a compelling depth effect, making them visually uninteresting. Moreover, directing such videos becomes challenging as viewers can look in any direction, diminishing the directive control from the content creator. Furthermore, higher FOV requirements necessitate increased resolution to meet acceptable Pixel-Per-Degree (PPD) resolution, leading to elevated hardware costs. This is because higher FOV videos are mapped into larger surfaces, decreasing their perceived resolution.

1.2 Goals

This project endeavors to surmount these challenges by:

1. Developing an Open-Source, Cost-Efficient, and Modular Stereo Video Camera Hardware System:
 - Introducing a hardware solution that is accessible to a broader audience, mitigating the cost barrier associated with existing options.
2. Implementing a Pre-processing Pipeline:
 - Mixing two 2D videos into the correct stereoscopic Side-By-Side (SBS) format to achieve an authentic depth effect.
 - Synchronizing audio files with the correct channels to create an immersive surround soundscape.
 - Incorporating metadata tagging through object and scene detection, facilitating streamlined browsing and organization of content.
3. Creating Intuitive VR Software for Content Management:
 - Designing software that enables seamless file browsing and content viewing within the VR environment.

This project not only addresses the critical gaps in hardware accessibility and the pre-processing pipeline but also represents a significant enhancement to existing methodologies. By fostering an open-source, modular, and cost-effective approach, this initiative strives to democratize VR content creation, making it more widely accessible and fostering innovation in the field.

Chapter 2

Background and Report of Literature Search

2.1 Lifelogging: A Methodical Approach to Memory Preservation

Lifelogging, a contemporary term encapsulating habitual documentation akin to social media practices, distinguishes itself through its methodical and routine nature. The motivation for lifelogging extends beyond sporadic capturing of moments to a deliberate effort to systematically record and preserve lifetime memories. A poignant illustration of this concept is evident in the Black Mirror episode that explores the immersive preservation of memories in an eye-camera format. This format aligns with human visual perception, eliminating the necessity for constant 3 Degrees of Freedom (3DoF) as our gaze isn't consistently omnidirectional. Although a broader Field of View (FOV) would enhance the lifelogging experience, it remains cost-prohibitive at present (human FOV approximately 220 degrees [1]).

2.2 The Significance of Virtual Reality (VR)

Virtual Reality (VR) transcends the conventional by offering spatial computing, a 3D environment, and an immersive emulation of real life. The emotional attachment fostered by VR content stems from its heightened realism, creating a sense of physical presence within the virtual space. This emotional resonance distinguishes

VR content from traditional media, providing a compelling reason for its adoption in content creation.

2.3 Windowed Viewing: Balancing Realism and Feasibility

Choosing windowed viewing over panoramic alternatives (180/360 degrees) acknowledges the balance between realism and current technological constraints. While panoramic views offer enhanced immersion, limitations in hardware capabilities, especially in the context of Free-Viewpoint Video (FVV) and 6 Degrees of Freedom (6DoF), necessitate a pragmatic approach. Windowed viewing, with its simplicity and affordability, aligns with the current state of VR hardware, ensuring a feasible and accessible content creation process.

Contrary to fully immersive FVV or reconstruction, the adoption of SBS format is grounded in the current limitations of VR Head-Mounted Display (HMD) hardware. While FVV offers unparalleled immersion and depth perception through 6DoF, practical constraints dictate a compromise. This decision is further supported by the spatial video capabilities of the iPhone 15 Pro, which, despite being 1080p at 60fps, aligns with the current standards for windowed style viewing. The emphasis here is not on resolution but on synchronized 60fps for a seamless experience.

2.4 VR Software for Content Browsing: Maximizing Interaction and Accessibility

Integrating VR software for content browsing amplifies the lifelogging experience. Leveraging the full potential of 6DoF through innovative UI design, expanded screen space, and interactive features such as timeline shelves, this approach redefines how users engage with their memories. The incorporation of hand tracking and eye tracking further enhances the intuitive and immersive nature of content navigation within the VR environment **tom-reference**.

The combination of lifelogging and VR represents a novel and niche subset within both domains. As technology advances and user preferences evolve, this integrated

approach is poised to become more mainstream, particularly in the realm of personal videos, memories, and photos. The synergy between lifelogging and VR anticipates a future where individuals seamlessly capture, relive, and share their most cherished moments in a more immersive and engaging manner.

2.5 Motion Sickness in VR

Persistences, resolution, eye strain, etc.

2.6 Examples of Existing Technologies for Lifelogging

Spectacles design: Snapchat Spectacles and Meta Ray-Ban glasses, but the latter is not stereoscopic. Chest mount/clip design: Narrative Clip, Insta360, and action cameras like GoPro. The most recent and striking one, however, is Apple iPhone 15 Pro, which uses its onboard main camera and wide lens camera together with machine processing to create convincing depth-accurate videos and photos at 1080p60fps. This is obviously done to allow users to create their content to be replayed back on their upcoming Apple Vision Pro XR HMD **iphone-pro-15**.

Chapter 3

Report on Technical Progress

The project is divided into 3 distinct parts, following the MoSCoW requirement framework (M: Must-Haves, S: Should-haves, C: Could-have, W: Wont-have).

3.1 MoSCoW Requirement

3.1.1 Hardware Development

- **M:** Develop an open-source, modular stereo video camera system with Raspberry Pi Pico microcontroller. Ensure it is low-cost and easily accessible.
- **S:** Consider additional features or improvements based on feasibility, such as using an onboard rechargeable Li-Po battery circuit instead of a power bank to power it.
- **C:** Explore advanced features like wireless connectivity or additional sensor integration, time permitting and if resources allow.
- **W:** Exclude features or components that are deemed impractical or beyond the scope of the project, such as higher resolution or different video format (180/360).

3.1.2 Software Preprocessing of Video

- **M:** Implement a pre-processing pipeline for transforming mono stills and video into stereoscopic Side-By-Side (SBS) format. Synchronize audio files for an immersive surround soundscape.
- **S:** Explore additional preprocessing features, such as metadata tagging through object and scene detection using existing library and tools.
- **C:** Automated video stabilization or advanced filtering options, based on available resources and time constraints.
- **W:** Exclude overly complex preprocessing tasks that may hinder the project timeline or exceed available resources, such as 3D depth reconstruction.

3.1.3 VR Software Application

- **M:** Develop an intuitive VR software application for seamless file browsing and content viewing. Ensure compatibility with the stereo video format.
- **S:** Implement innovative UI designs for enhanced interaction within the VR environment.
- **C:** Explore the integration of hand tracking, and eye tracking, if resources and time allow.
- **W:** Exclude overly ambitious features that may compromise the core functionality or extend the project beyond feasible timelines, such as a personal AI assistant.

3.2 Hardware System

3.2.1 Cost Analysis

The budget provided for the project is £150, and the aim is to get all components under £100, making it a low-cost solution compared to alternatives that usually cost around £300 or more. The most expensive part, as expected, is the camera modules, which are around £35 each. However, they boast a 5MP sensor and are capable of taking 1080p60fps video, justifying their cost. The exact component details and costs can be seen in Appendix 1.

3.2.2 Microcontroller and Ecosystem

After comparing costs, availability, ease of use, and hardware constraints, the choice of microcontroller is Raspberry Pi Pico due to its cheap cost, cohesive documentation, compatible hardware, and, most importantly, buffer size, which is important to get high enough resolution images/video to prevent motion sickness.

To achieve an immersive experience, audio is also an important variable. Thus, the use of 2 independent electret microphone modules is added to work in tandem with the camera. This, in theory, should make it possible to achieve stereo sound channels for each ear.

An SD card extension board is also added to host the SD card that holds all the data. The circuit schematics can be seen in [Figure 3.1](#)

3.2.3 Constraints

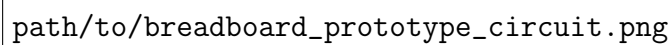
The initial draft design is to have both cameras mounted on the side of eyeglasses, akin to Ray-Ban Meta and Snapchat Spectacles glasses. However, after getting the camera module and other components, it is deemed too unwieldy and difficult to fit the electronics into a small constraints form factor. The main reason for this choice is to capture the footage as close as how humans see the world, and mounting the camera close to the eye would achieve that compared to a chest mount design.

To compromise, a hat/cap-mounted design is chosen, where the POV is higher than usual, but the camera movement/rotation will still follow head movement and should give a realistic enough POV as seen in [Figure 3.2](#).

Trying to make it power-efficient might prove challenging and needing to deal with additional circuitry, so for now, the system will be powered with a power bank to the micro-USB on the main Pico board.

3.2.4 Onboard Embedded Software Algorithm

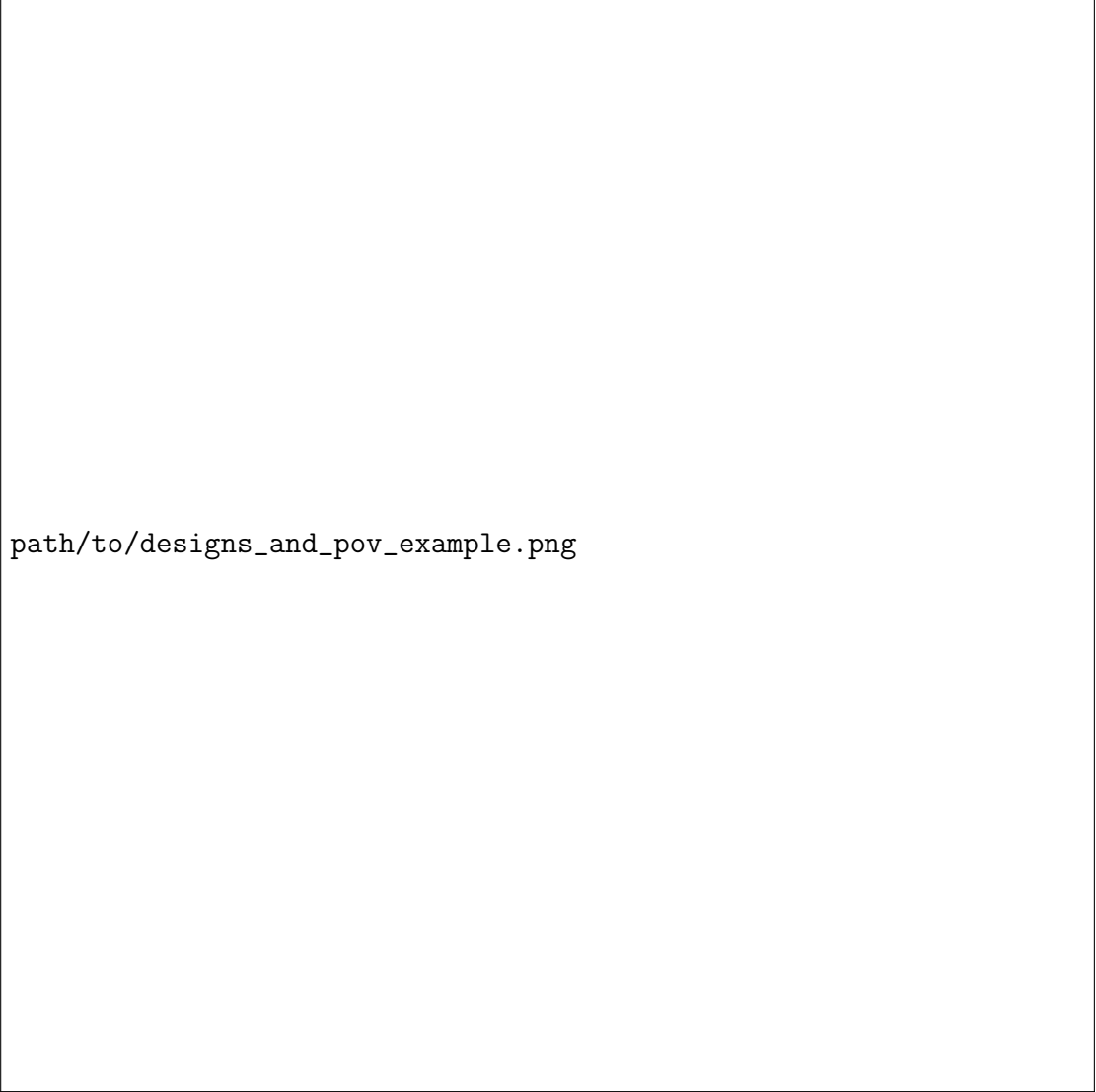
The main idea is that in normal lifelogging mode, the cameras would take stereo still images every now and then on a timed interval, which will be decided through trial and error and optimization depending on how much storage the images occupy and the power efficiency of the algorithms. However, to get a more immersive

A large rectangular box containing the text 'path/to/breadboard_prototype_circuit.png', which serves as a placeholder for the circuit diagram.

path/to/breadboard_prototype_circuit.png

FIGURE 3.1: Breadboard Prototype Circuit

experience, video is also needed, and a button can be used to manually start and stop recording, with an LED being an indicator when it's recording. The saved files should be aptly named for easier processing later, maybe in a standardized DATETIME-NUMBER format. The proposed algorithm can be seen as in Figure [3.3](#)



path/to/designs_and_pov_example.png

FIGURE 3.2: Designs and POV Raycast Example

3.3 Video Pre-Processing Software

3.3.1 Motivation

The motivation behind implementing a pipelined process for video pre-processing stems from the acknowledgment that attempting to execute this intricate task directly on the Raspberry Pi Pico would introduce unnecessary complications and exceed its memory and processing power limitations. By adopting a pipelined approach, the pre-processing burden is shifted to a more capable system, ensuring a more efficient and effective transformation of videos and images. This is particularly crucial for the generation of Side-By-Side (SBS)/3D format from two



FIGURE 3.3: Flow Chart for Onboard Embedded Software Algorithm

mono/2D stills using ffmpeg. While ffmpeg provides a baseline solution, the desire for a streamlined script or process is imperative to facilitate batch processing, ensuring accuracy and effectiveness. Additionally, the inclusion of a custom metadata tagging system for object and scene detection is pivotal. This customized metadata enhances the overall VR experience by enabling advanced filtering and indexing, thereby optimizing the video and image browsing experience within the VR application.

3.3.2 FFMPEG for Mono to Stereo Stitching

The utilization of FFMPEG for mono to stereo stitching serves as a cornerstone in the pre-processing pipeline. FFMPEG, a powerful multimedia processing tool,

efficiently transforms mono/2D videos and images into the desired stereoscopic Side-By-Side (SBS)/3D format. This process is instrumental in creating a lifelike and immersive visual experience for VR content, aligning with the project's goal of democratizing VR content creation.

3.3.3 Object and Scene Detection

Incorporating object and scene detection algorithms further enriches the preprocessing pipeline. Leveraging advanced computer vision techniques, this step aims to automatically identify and tag objects and scenes within the videos and images. The integration of object and scene detection not only enhances the visual content but also lays the foundation for sophisticated filtering and indexing capabilities within the VR application. This ensures that users can seamlessly navigate and explore their lifelogging content with enhanced precision and relevance.

3.3.4 Custom Metadata Tagging

Custom metadata tagging is a pivotal aspect of the preprocessing pipeline, allowing for the incorporation of user-defined information related to object and scene detection. This bespoke metadata adds a layer of personalization to the content, enabling users to categorize and organize their lifelogging data according to individual preferences. The inclusion of custom metadata serves as a cornerstone for optimizing the VR content browsing experience, ensuring that users can easily locate and revisit specific moments within their immersive collection.

3.4 Video Player VR App

3.4.1 Game Engine – Godot 4.1

The selection of the Godot 4.1 game engine for the development of the video player VR app is rooted in the project's commitment to Free and Open Source Software (FOSS) principles. Unlike more established game engines such as Unity or Unreal Engine, Godot aligns with the FOSS ethos, making it the ideal choice for this project. Despite having fewer documentations and examples compared to its counterparts, this presented an opportunity for active participation in its development.

Consultation with the Godot community, particularly through their Discord channel, guided the decision-making process, revealing that utilizing shaders is the most straightforward approach for rendering stereo Side-By-Side (SBS) video playback.

3.4.2 Shaders

The implementation of shaders in the Godot 4.1 engine for stereo video playback is relatively uncomplicated. Given the SBS format of the video, the left-half is rendered for the left camera eye, and vice versa for the right-half. The `gdshaders` code utilized for this purpose adheres to this logic, facilitating an efficient rendering process.

3.4.3 User Interactions (UI)

The user interface (UI) for the video player VR app is designed within a 3D space, a standard practice for VR applications and games. However, a challenge arises as the video player itself operates as a 2D screen within this 3D environment. Initial attempts to follow a tutorial by MalcolmNixon on YouTube, supplemented by consultation with the Godot community, encountered difficulties in integrating the shader script within the same scene as the `videostreamplayer` node. Subsequent experimentation and development efforts proved inconclusive.

Fortunately, after seeking guidance from the Discord community, MalcolmNixon, the tutorial's author, provided a pivotal solution. Instead of employing the `2DScreenIn3D` approach, the revised method involves instantiating a 2D screen `videostreamplayer` node in the main scene, where the shader is then applied. This modification successfully addresses the challenges encountered during the development process. Example of app running is seen [Figure 3.4](#)

3.4.4 File Browsing

While file browsing functionality has not been fully implemented, conceptualization has begun. Proposed ideas include leveraging metadata tagging for specific searches and employing bookshelves or 3D objects as interfaces for navigating between months or weeks, deviating from the conventional 2D screen approach to enhance interactivity.



FIGURE 3.4: Images of App Running

To optimize browsing efficiency, the screen space may be expanded to resemble an ultra-wide monitor or more, enabling users to view a greater number of files in a single window. This approach capitalizes on the immersive capabilities of VR, utilizing the 360-degree view and 6 Degrees of Freedom (6DoF). Additionally, the implementation of the depth axis remains contingent on contextual considerations and future developments.

Bibliography

- [1] C. Gurrin, A. F. Smeaton, and A. R. Doherty, *LifeLogging: Personal big data*, 2014. DOI: [10.1561/15000000033](https://doi.org/10.1561/15000000033).