Electronics and Computer Science

Faculty of Engineering and Physical Sciences

University of Southampton

Muhammad Hazimi Bin Yusri

24/11/2023

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

Project supervisor: Tom Blount

Second examiner: <2nd Examiner>

A project progress report submitted for the award of

<MEng Electrical and Electronics Engineering>

Abstract

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 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.

Abstract should be written last according to what u wrote to reflect rest of document, progress report abstract is different than end project report abstract.

Typical structure is like inverted pyramid)

- Introduction (problem statement, scope, goals)

- Background of existing technologies (can separate/structure however I want), and its considerations (where most reference comes from are research, academic papers)

dont need to be exactly like what im doing, separate it for example (lifelogging and virtual reality), and talk about how ur design is combination of both on next section

- Design, who its for, what, why? (Can be moscow requirement)

Circuit diagram, codes when its existing in report, acquisition, budget and cost analysis.

Make sure all diagrams/pictures are related to paragraph its in. Dont need to be my images/pictures for reference, but make sure you got the rights for it.

(usually larger and more important one)

- Project management, how u manage time (gantt chart), risk assessment, how do u plan of next half of project.

max 3000 words (not including references)

if enough space

-conclusion of what i do and what i do next

content are more important than heading names

2 Background and report of literature search

Why lifelogging?

Just old terms for current habits of social media, only difference is lifelogging is more methodical and routine. Also bring a good point for comparison instead of just randomly taking videos/photos every now and then. Inspiration from black mirror episode to ‘preserve’ and record lifetime memories, especially in immersive mode like SBS, just how like we see the world, thus no need 6DoF as we are not looking around us 360 all the time, having better FOV (human fov around 220? Need cite), would be cool but too expensive for now.

Why VR?

VR is virtual reality, spatial computing, space, 3D, emulating real life etc.

More emotional attachment as the content is more immersive/real, feels like you are there physically.

Why windowed instead of 180/380?

Why SBS instead of FVV/reconstruction?

Limitations of current VR HMD hardware for real FVV, which is more immersive and uses 6DoF and have depth but current hardware for VR themselves is not good enough yet, so basic SBS is sufficient enough Affordable content creation for free-viewpoint video and VR/AR applications - ScienceDirect, 6.2.1. data format limitations.

This also can be seen by iphone 15 pro adoption of spatial video, which is only 1080p 60fps, which prove 1080p is good enough for now for window style viewing. Only problem is I need good synced 60fps.

Why VR software to browse content?

Can use the 6DoF with innovative UI design to its fullest, can have more screen space for loads of content, have a shelf of timeline etc, so many possibilities. (cite needed, one of paper tom emailed) + can use hand tracking for more intuitive

++ use eye tracking.

Why combine both?

Novel and niche subset of both, but surely will become more adopted and mainstream in future, especially for personal videos/memories/photos etc.

Contents

[Abstract 2](#_Toc151382824)

[Contents 3](#_Toc151382825)

[1 Introduction 4](#_Toc151382826)

[2 Background and report of literature search 5](#_Toc151382827)

[3 Report on Technical Progress 6](#_Toc151382828)

[4 Plan of remaining work 7](#_Toc151382829)

[References 8](#_Toc151382830)

# Introduction

## 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.

## 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.

# Background and report of literature search

## 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 [cite needed]).

## 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.

## 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 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.

## Side-By-Side (SBS) Format: Overcoming Hardware Limitations

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.

## 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 (citation needed, reference provided by Tom).

## Synergy of Lifelogging and VR: A Niche for Future Mainstream Adoption

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.

## Motion Sickness in VR

Persistences, resolution, eye strain etc.

## 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. 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 own content to be replayed back on their upcoming Apple Vision Pro XR HMD.

# Report on Technical Progress

The initial design is to have it mounted on top of a cap, providing a more realistic point-of-view (POV) compared to chest-level strap design, but not constrained with too small of form factor like spectacles design.

IMAGES OF DESIGN 1,2,3 (CHEST, TOP CAP, SPECTACLES)

Main design is aimed towards lifelogger-esque/timelapse algorithm where pictures are taken at timed intervals.

There will be buttons for manual recording with LED light indicator.

The video also includes stereo audio from two individual electret microphone.

The project is divided to 3 distinct part as per following MoSCoW requirement framework (Must-Haves, Should-haves, Could-have, Wont-have).

## Moscow requirement

### Hardware Development:

* **Must-haves:** Develop an open-source, modular stereo video camera system with Raspberry Pi Pico microcontroller. Ensure it is low-cost and easily accessible.
* **Should-haves:** Consider additional features or improvements based on feasibility, such as enhancing the Field of View (FOV) or exploring power-efficient components.
* **Could-haves:** Explore advanced features like wireless connectivity or additional sensor integration, time permitting and if resources allow.
* **Won't-haves:** Exclude features or components that are deemed impractical or beyond the scope of the project.

### Software Preprocessing of Video

* **Must-haves:** Implement a pre-processing pipeline for transforming mono videos into stereoscopic Side-By-Side (SBS) format. Synchronize audio files for an immersive surround soundscape. Incorporate metadata tagging through object and scene detection.
* **Should-haves:** Explore additional preprocessing features, such as automated video stabilization or advanced filtering options, based on available resources and time constraints.
* **Could-haves:** Investigate machine learning algorithms for more sophisticated scene and object detection, if time and resources permit.
* **Won't-haves:** Exclude overly complex preprocessing tasks that may hinder the project timeline or exceed available resources.

### VR Software Application:

* **Must-haves:** Develop an intuitive VR software application for seamless file browsing and content viewing. Ensure compatibility with the stereo video format.
* **Should-haves:** Implement innovative UI designs, hand tracking, and eye tracking for enhanced interaction within the VR environment. Consider additional features like content organization tools or social sharing functionalities.
* **Could-haves:** Explore the integration of AI-driven features, like personalized content recommendations, if resources and time allow.
* **Won't-haves:** Exclude overly ambitious features that may compromise the core functionality or extend the project beyond feasible timelines.

## Hardware System

### Cost Analysis

The budget provided for the project is £150, I aim to get all components under £100 to make it low-cost compared to other solutions that usually goes around £300 or more. The most expensive part, as expected are the camera modules, which are around £35 each, however they boast a 5MP sensor and capable of taking 1080p60fps video so it is worth the cost. The exact components details, and cost can be seen at Appendix 1.

### Microcontroller and Ecosystem

After comparing costs, availability, ease of use and hardware constraint, 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 immersive experience, audio is also an important variable, thus the use of 2 independent electret microphone module is added to work in tandem with the camera. This in theory should make it possible to achieve stereo sound channels for each ear.

SD card extension board is also added to host the SD card that holds all the data.

### Constraints

The initial draft design is to have both camera mounted on the side of eye-glasses, 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 small constraints form factor. The main reason for this choice is to capture the footage as close as how human sees the world, and mounting the camera close to eye would achieve that compared to chest mount design.

To compromise, hat/cap mounter 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.

IMAGES OF 3 DESIGN AND THEIR POV RAYCAST EXAMPLE

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 micro-USB on the main Pico board.

### Onboard Embedded Software Algorithm

The main idea is that on 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 optimisation depending on how much storage the images occupy and power efficiency of the algorithms. However, to get more immersive experience, video is also needed, and button can be used to manually start and stop recording, with having LED being indicator when it’s recording. The saved files should be aptly named for easier processing later, maybe a standardised DATETIME-NUMBER format.

IMAGES/DIAGRAM OF FLOW CHART FOR ALGORITHM

## Video Pre-Processing Software

### 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 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.

### 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.

### Object and Scene Detection

Incorporating object and scene detection algorithms further enriches the pre-processing 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.

### 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.

## Video Player VR App

### Game Engine – Godot 4.1

Why Godot instead of other established game engine like Unity or Unreal Engine? Main reason is to complement the spirit of Free and Open Source Software (FOSS) of this project so Godot which is another FOSS is the perfect choice. It had less documentations and examples compared to the other two but this is also perfect chance for me to help spearhead the development and be the one that helps implement it. First, after consulting the devs in their discord channel on the easiest way to render the stereo SBS video playback, it seems using shaders are the easiest way.

### Shaders

Pretty straightforward apparently, as the video is SBS format, left-half of video is rendered for left camera eye and vice versa for right-half. Here is the gdshaders code used:

### User Interactions (UI)

The app loads user into 3D space as usually done for VR applications and game, thus UI

for everything will also be in 3D. But the video player is just a 2D screen, so how?

At first, I followed one of the tutorial using 2DScreenIn3D video player tutorial by MalcolmNixon in Youtube and consulted Olej in Discord, but after experimenting and developing for a while, I was stuck as the shader script need to be in same scene as videostreamplayer and should be in main scene, and I was unable to resolve this problem.

Thankfully, after asking for guidance in Discord again, the author of the tutorial, MalcolmNixon comes for a clutch help by providing me with example of stereo video player which instead of using 2DScreenIn3D, just instantiate a 2D screen videostreamplayer node in main scene, and use the shader there. It works perfectly.

### File Browsing

It has not been implemented yet, but few ideas so far is to make use of metadata tagging for specific searches, and using bookshelves/3D object as interface for going between months/week instead of usual 2D screen to increase interactivity.   
  
For quick and efficient browsing, the screen space can also be widen to be like ultra-wide monitor or more so user can see more files in one window making use of 360 degree and 6DoF of VR. Another axis, the depth axis may also be implemented depending on circumstances.

# Plan of remaining work

12 pt text for main body

References