UNIVERSITY OF CAPE TOWN

MASTER'S THESIS

Investigating the virtual directing strategies of a Virtual Cinematographer in an Automatic Lecture Video Post-Processing System

Author:
Mohamed Tanweer KHATIEB

Supervisor:
Patrick MARAIS & Stephen MARQUARD

A thesis submitted in fulfillment of the requirements for the degree of Master's in Computer Science

June 20, 2023

Declaration of Authorship

I, Mohamed Tanweer KHATIEB, declare that this thesis titled, "Investigating the virtual directing strategies of a Virtual Cinematographer in an Automatic Lecture Video Post-Processing System" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Signed by candidate

Date:

UNIVERSITY OF CAPE TOWN

Abstract

Faculty of Science

Department of Computer Science

Master's in Computer Science

Investigating the virtual directing strategies of a Virtual Cinematographer in an Automatic Lecture Video Post-Processing System

by Mohamed Tanweer KHATIEB

As recording technology improves and becomes more affordable, many learning institutions are using lecture recording to make lessons more persistent and accessible. Statically mounted 4K cameras are now cheaper than PTZ cameras which makes them a desirable alternative for lecture recordings. Unfortunately, 4K resolution videos are very large, posing a problem for storage and streaming - the file size for a 45 - 60 minute lecture video in 4K can exceed 2GB. Many students cannot afford the bandwidth required to stream such large files. Furthermore, since static 4K cameras do not move, they require a wide-angle view of the venue in order to capture as much of the front of the venue as possible. This view is much too zoomed out for viewers to see the details, such as writing on the boards and the presenter's facial expressions, captured by the 4K resolution.

This dissertation investigates an approach to post-processing these 4K lecture videos to reduce the file size and emphasise lecture details such as lecture motion and board/screen usage. This is done using scene tracking data (generated via a third-party front-end) which a Virtual Cinematographer (VC) uses to make decisions on about which areas to crop from each 4K frame in the original video. The VC then positions and sizes the cropping windows in such a way that the resultant, cropped video resembles one recorded by a human camera operator. This is accomplished using cinematographic heuristics to inform its decision-making.

The VC uses scene analysis algorithms to determine how the environment changes as time progresses in the video. By dividing the video into "chunks" (equivalent to "scenes" in traditional cinematography) based on context, the VC is able to maintain stable shots with consistent framing to avoid jittery and disorienting footage. These contextual chunks are determined by comparing the trajectory of the presenter with the manner in which the features on the board regions change over time. After the chunks are established, the VC creates transitions between them while avoiding any changes to the framing inside each chunk. The final output is a JSON file containing the cropping coordinates for each frame in the video for a third-party video cropping application to use when producing the final video.

We performed a user evaluation of the VC to measure user satisfaction with the resulting output videos and how successful it was at following its heuristics. The VC succeeded in following the major heuristics such that viewers were satisfied with the output based on the framing of the presenter and the content on the boards, transition stability and smoothness of motion, and transition frequency with the VC only changing shots when necessary.

Acknowledgements

I begin by thanking the Almighty for everything and for guiding me every step of the way.

This work would not have been possible without the support and guidance of my supervisors Patrick Marais and Stephen Marquard. Their patience and insight with my many iterations have been invaluable in shaping this dissertation and communicating the content as clearly and effectively as possible.

A big thank you to my friend Charles Fitzhenry for his work on the front-end tracking module used as the input for my Virtual Cinematographer as well as his assistance and support throughout the writing of this dissertation.

I would also like to thank my friends and fellow masters students for the roles they played in this journey.

My family has been patient with me throughout my writing this dissertation and I wish to thank them for their support and understanding.

Finally, it is important to acknowledge and thank the University of Cape Town, especially the Department of Computer Science, for providing the required resources and facilities to complete this dissertation.

Thank you to everyone who had a part in helping me to complete this research endeavour and making it worthwhile.

Contents

Chapter 1

Introduction

Rapid technological advances are enabling learning institutions to explore new approaches to conveying learning material to students. One such approach is Automatic Lecture Recording (ALR) and the accompanying playback systems. Lecturers need only present and record one physical lecture and students can then review the content as many times as needed to grasp the lesson. Not only is it possible for such learning systems to be integrated with IT infrastructure at higher education institutions, but the technology is becoming cheaper. 4K cameras (cameras recording footage at 3840×2160 pixels), which are positioned and focused to have a wide-angle view, are able to capture more detail in the venue than Pan Tilt Zoom cameras (with a similar zoom and focus) and older generation cameras.

The file size of a 45-minute lecture video recorded at 4K can exceed 2GB, which is too large for affordable and practical streaming. Such large files cannot be stored indefinitely since the number of lectures recorded daily will eventually overwhelm the hosting platform's storage capacity. Furthermore, it may not be feasible for students to stream a large number of these video files due to budget and network constraints. These problems are exacerbated in developing countries which do not have access to cheap, reliable, high-bandwidth internet connections. Using a lower-resolution camera or reducing the resolution of the output video files does not solve this problem, however, as the wide-angled view includes the whole front region of the classroom. Under such capture conditions, the video appears 'zoomed out' and the content on the boards, as well as the presenter's facial expressions, are too small to see comfortably. A more advanced technique is required to solve both issues simultaneously.

The solution must reduce the file size whilst allowing the viewer to see close-up views of the presenter and the boards.

This could be done by cropping regions out of the original 4K frames such that the less relevant content is removed from view. Under this approach, the lecturer and the boards must be tracked in order to determine which areas of the frame should be included or cropped out. Unfortunately implementing this approach naively introduces a complication: since video frames are presented sequentially in time, differences in the position/offset of the cropping window across a sequence of frames will generate jittery video, which is visually jarring. The footage may also appear distorted since the cropping regions will often change size according to what the context tracking algorithm selects as important. The solution should, therefore, utilise guidelines on how to avoid these complications and make the video as comfortable to view as possible.

1.1 Research Aims

This work aims to create a system that uses tracking information about the content of a lecture video to crop out the regions of importance using cinematography heuristics (guidelines). The purpose of these heuristics is to prevent jittery and distorted output footage. They should also smooth the trajectory of the cropping window across video frames so that the output video resembles a video recorded by a human camera operator. The resulting system is referred to as a 'Virtual Cinematographer' (VC) and operates on pre-recorded lecture videos in post-production.

The system relies on the output of a tracking front-end that uses temporal differencing to detect motion and writing detection algorithms to identify board regions and how they change. The VC processes this tracking information and generates a JSON file containing the coordinates for each cropping window in the larger 4K frames which is then used by an external application (such as FFMPEG) to create the final output video. This is an area of ongoing research and our heuristics build on what has been done in previous studies, which we discuss in more detail in Chapter 2.

The VC's core heuristics are the following:

- The VC must only show what is relevant and important to the viewer
- The VC must only perform transitions when necessary
- Transitions should have accelerations and decelerations to create smooth movements

There are other heuristics which the VC uses, but these heuristics are the minimum requirements for the VC's success.

1.2 VC Framework

This section gives an overview of the VC module's role in a larger system in terms of its input, processing and output stages.

Input The VC reads the data from the input JSON file and creates an internal representation in memory (to reduce disk access inefficiencies). The JSON file is no longer required after this step in the program as all the classes and data structures are populated with the necessary information.

Processing The VC identifies the presenter and any writing regions for each frame. Scene analysis helps to determine the motion of the presenter and the changes in the writing on the boards (using the tracking information) and this guides the VC's decision-making process on how best to position the cropping window for each frame in the video.

Ouptut The VC produces a JSON file containing the coordinates for each frame's cropping window. This file is then used by an external application to write the final video file.

1.3 Contributions

This research contributes to the field of Automatic Lecture Recording by introducing a post-processing Virtual Cinematographer that crops regions from the frames in a 4K resolution video rather than driving a motorised Pan Tilt Zoom (PTZ) camera. The VC operates in post-production because real-time solutions only have access to past and present information whereas a post-processing solution can also access information from frames ahead of the current one (also referred to as 'future information').

The cameras used to record the 4K footage are statically mounted and do not move around. They are focused on a wide-angle view of the front area in the lecture venues and the recorded area does not change. With this view of the venue, the VC is able to perform pan and tilt operations, while simultaneously zooming in or out as required. This is because the field of view remains static and the VC controls the cropping window's size and position for each frame.

The heuristic algorithms employed by the VC are the main contribution of this dissertation since they instruct the VC on where to place the cropping window coordinates and how the final video should appear to viewers. The primary heuristics employed by the VC are guiding the viewer's attention to what is most important (and relevant), keeping camera transitions smooth by adding acceleration and deceleration and limiting the number of transitions by only adjusting the camera's framing when necessary.

These heuristics are programmed using scene analysis algorithms based on the tracking information provided by the front end of the system.

The user evaluation provides insight into the success of the VC and it indicates where the VC can be improved in future iterations. Our survey of the literature did not find many investigations into the use of wide-angle 4K footage being cropped in post-production in a similar way to how the VC operates, which makes the results from this evaluation useful for future projects exploring similar topics (or improving on this work).

The code used to program the Virtual Cinematographer was written using C++ and the OpenCV library. The VC is open-source and the code will be released at the end of the study.

1.4 Limitations

The primary limitation of the Virtual Cinematographer is its dependency on the tracking information provided by the tracking front end since the VC does not deal with the video-tracking aspect of the system in this research.

The focus of this investigation is making a module for a larger system that is specialised in cropping video files using cinematographic heuristics. This dependency limits how the VC's heuristic algorithms are written because a limitation in the front end is a limitation of the VC. For example, the VC does not have the option to favour framing the boards over the presenter. This feature would allow the VC to focus on the boards in cases where the presenter paces a lot without adding anything new to the boards (or projector screens). This focus-switching operation requires gesture and face tracking, which could not be reliably obtained by the available tracking front-end due to its limitations.

Another limitation of the VC is that it does not create the output video files. It writes the cropping window coordinates to a JSON file for third-party software to use when cropping the original video into the final output.

1.5 Thesis Structure

The remainder of this thesis is structured as follows:

- Chapter 2 introduces the field of Automatic Lecture Recording and covers a brief system breakdown before mentioning the history of the field and it concludes with a brief explanation of the core statistics used in this study and reviews the literature in the field of ALR in terms of real-time processing, post-processing, camera setup, and framing heuristics
- Chapter 3 covers the framework of the Virtual Cinematographer and goes into detail about the system input, processes, and output
- Chapter 4 discusses the methodology of the user evaluation and goes into detail about the experiment design
- Chapter 5 presents the results of the user evaluation and discusses the findings in the context of the evaluation and the VC's operations
- Chapter 6 concludes this thesis and highlights the potential for future research

Chapter 2

Background and Related Work

This chapter provides an overview of the technical aspects of lecture recording in order to present the background necessary to understand our contribution. A brief overview of the history of lecture recording is also presented.

2.1 The Field of Automatic Lecture Recording

Lecture Recording is the practice of recording lectures and making the lecture material available online such that students are able to access this content outside of the lecture venue. It allows students to review the lecture at a later time and gives absent students the chance to catch up on the missed material. Recent advancements in camera quality, and reductions in cost, have caused many institutions to adopt the use of lecture recording as a supplementary form of conveying material to their students [36]. Since the beginning of lecture recording for the purpose of remote learning, research has been done to automate the recording process. This gave rise to a new field, namely, Automatic Lecture Recording (ALR) [18].

ALR systems do more than merely record the venue as the lecturer conveys the material by placing a wall-mounted or ceiling-mounted camera adjusted to a single field of view (known as a static camera). Some control is needed to better manage what is shown by the camera in a given context [19]. This requires careful thought from the developers since the way a lecture is portrayed is very different from the way a film or television series is recorded. Some guidelines on the best ways to record lectures are explained in some of the literature of the field [16], [19], [21].

2.2 System breakdown of Automatic Lecture Recording

There are multiple disciplines and technologies involved with the field of ALR (e.g. camera operation requires knowledge of videography, computer science is required to create an automated camera operator and to handle any networking for the recorded videos) [29]. The field has grown in popularity around the world in teaching institutions as a reliable approach to incorporate remote learning [21]. We refer to the following works in this section when describing the typical system breakdown of an ALR system [32], [33], [39].

The process of ALR can be divided into the following components:

1. Input system
2. Storage and networking of data
3. Post-processing of data
4. Output system

2.2.1 Input System

Most automated approaches to recording lectures use one of two input devices:

Pan Tilt Zoom Cameras: These devices (also called PTZ cameras) are able to move on 3 degrees of freedom using onboard hardware for automatic local control or remote manual control. It can pan across the venue, tilt up or down and zoom in or out in order to get a better framing of the content.
We refer to these works when talking about Pan Tilt Zoom cameras [38], [43], [51].

We do not focus on using PTZ cameras in our research since they require real-time control which limits the temporal scope for making decisions in a given context.

Static Cameras: These devices use a high-resolution camera which remains fixed to the wall or ceiling such that it cannot move. The focus of the camera for this assembly would depend on the layout of the venue in which the camera is installed. Once the camera is set up in a venue, it only needs to record and communicate its data without requiring any feedback from the processing module. The data recorded in such a system needs to be communicated quickly and completely since the subsequent steps in the system depend on its output.

Our research focuses on an ALR System which uses a Static Camera setup for recording videos. These cameras are given a wide-angled view such that the entire front region of the lecture room is visible. This view is then reduced in post-production which we discuss further in the post-processing section.

2.2.2 Storage and Networking of Data

The Storage requirements of the video footage depend on the resolution of the video footage being recorded, the speed (or multi-tasking capabilities) of the processing module, and the budget allocated to the recording system. It is in the best interest of the institution implementing an ALR system to ensure the system being implemented does not fall short on storage since later modules require the footage of the video to accomplish their assigned tasks.

Solutions which employ a PTZ camera need to have some way to drive the movements of the camera. In modern versions of such a system, the processing is done either in the camera module or by a local extension to the camera module and the information does not require any network transportation except for storing the lecture recording on a server.

Solutions which employ static cameras only need to send data to the server without having to wait for instructions to come back from the server.

In either case, the network infrastructure needs to be of high quality. Data sent from the camera must be sent to the server as fast as possible and the information must not be lost due to dropped packets and other networking issues. In the event of a network failure, most ALR systems have a local Storage point so that lectures and lecture recording can continue unhindered without the lecture recordings being lost. The recordings which are stored locally are then uploaded as soon as the network is restored. There is a variant of this style where the system has a local server to queue and store the recordings (and sometimes process them locally as well) and data is only uploaded during times of low network traffic to avoid inconveniencing the viewers. Fast and complete data transference will allow the server to process the video optimally and (depending on the camera system implemented) will produce better instructions and output.

2.2.3 Post-Processing of Data

Scene Analysis is the term given to what the ALR system does in order to understand what is happening in the venue before it can decide what instructions it must send to the recording module (if any). This process makes use of many computer vision algorithms to gain an understanding of the scene.

The following are some of the techniques which are common in many implementations:

Background Subtraction involves comparing two subsequent frames directly by subtracting the pixel values at every position in the second frame from each of the corresponding pixel values in the first frame [47]. This provides the processing module with some insight as to how much movement has happened between the two frames.

Edge Detection allows the processing module to isolate one item in the scene from another by identifying the borders which separate them [46]. There are a few techniques which can be used in order to get this information, but all of them look for a sharp change in colour or contrast in order to identify an edge.

Clustering is the process of grouping points in close proximity to one another such that one grouping is separate from another grouping [56]. This is done by comparing the distance from every point to a seed point (an anchor point). The points which are closer to the seed point than a specified maximum distance get assigned to the cluster.

Feature Detection looks for features in a given matrix of pixels. Any sharp corners, changes in direction or line intersections are highlighted as features [53]. When this technique is paired with clustering it can help to identify specific entities in the scene (e.g. Lecturer, blackboard, desk, etc.).

These techniques are the building blocks for scene analysis and any action performed by the lecturer or event in a classroom can be constructed from these building blocks.

The Processing Module can operate in one of two modes: *Real-Time Processing* mode or *post-processing* mode.

Real-time processing is where the camera must make adjustments at the time of recording [50]. To achieve this the entire pipeline must be streamlined as much as possible. This could mean making the server (or processing module) in the same location as the camera such that there is no networking required or reducing the number of considerations during the processing step and making more informed guesses rather than confident decisions. It is also important to note that real-time processing means there cannot be much storage of data and the system will not require as much storage in their system (aside from the publication platform).

Post-Processing is the core of our study. The recorded video must be evaluated and, if not fit for publication, modified such that the video conveys the material covered in the lecture in the best way possible [33]. The type of camera used in a post-processing system is most likely a stationary one since the video is already recorded and there is no point in having remote control over the camera in such a situation. Post-Processing a recording would therefore involve the alteration of the frames in the video file to produce the desired effect.
In our case, we cropped the original frames such that only the most important information is shown to the viewer. This is done in such a way that the final video appears to have been recorded using a moving camera by adjusting the position of the cropping region incrementally as the video progresses so as not to disorient the viewer.

There are advantages to using post-processing rather than a real-time solution, namely, the processing module has more time in a post-processing solution and can therefore dedicate more resources to making better decisions on the best ways to crop the frames. A post-production implementation has access to the entire video and can therefore use a much larger temporal context by using past, present and future information when deciding what to do for a particular camera shot. An additional benefit to cropping the frames is that the output video will be smaller in size which makes streaming the video easier for viewers [15], [29].

This approach takes pre-recorded video footage out of storage and makes decisions on how to modify the data using a modelled approach [29]. There are usually mathematical formulae or heuristics guiding the decision-making. This approach is much more dependent on storage and a bigger budget should be allocated for storage in such a system.

Speed is less of an issue in a post-processing solution and so the output quality is expected to be much better as a result. In some cases, the processing module makes multiple passes over the data and its decisions in order to build up confidence or gain context at different scopes of the footage. The post-processing module is sometimes separated into smaller roles and each implementation has different names for these roles.

Virtual Camera Operator

The role of the Virtual Camera Operator (or just Virtual Camera) is to manage the recording of the video footage [30], [50]. In a PTZ setup that means physically moving the camera to a view. This means the server or the processing module must communicate with a Virtual Camera Operator to get the PTZ camera to adjust its view and focus. A Virtual Camera Operator in a setup with multiple static cameras in a real-time solution must record the footage and send this information to a Virtual Director and then either change the focus or priority of the camera assigned to it based on instructions given by the Virtual Director.

Virtual Director

The role of the Virtual director (VD) is more high-level than that of the Virtual Camera Operator. A VD must perform scene analysis and make decisions based on which camera feed (or area of the venue) shows the most important information and excludes the irrelevant things [16], [17]. To achieve this a VD makes use of mathematical functions which optimise particular behaviours or it uses heuristics which will guide its decision-making. Once a decision is made, the VD either sends instructions to a Virtual Camera Operator (in a real-time solution) or modifies the video file directly (in a post-processing environment).

Virtual Cinematographer

This is a combination of the VD and the Virtual Camera Operator. The role of the Virtual Cinematographer (VC) is to direct the camera's operations to show the viewers the most relevant information in the most aesthetically appealing way possible [19], [29], [44]. This is achieved by directly manipulating the view of the camera (in the case of a static camera this would be the cropping region rather than the actual camera). It makes deductions about the environment based on the information provided from scene analysis and decides how the camera should move to achieve the best framing. These decisions are guided by mathematical functions or heuristics which instruct it on how to adjust its view with the given context.

This research focuses on the VC module in the larger system of ALR and so there is more to be mentioned about the VC. Instead of moving a camera physically, our research focuses on using a static camera with a wide-angle view and then cropping every frame to reduce the field of view of the video such that it appears like the camera is moving. To make this less jarring and easy to follow for the viewer, the VC adjusts the speed of the cropping window to give the illusion of a camera pan, tilt or zoom. Another benefit to cropping the frames is that the size of the video file is reduced which helps with streaming the video.

2.2.4 Output System

The processed video footage is saved as a new video file and released for publication. The institution usually uses its own website to host the videos for viewers to consume [27].

Creating the output video varies on the type of camera setup and the way the Virtual Director has managed the footage from those cameras. Post-Processed PTZ systems can apply visual effects on top of the recorded video before the footage is saved to a new file and the same can be done for static cameras (if required) since these cameras are generally used in a post-processing approach [26], [34].

2.3 A Brief History of the Field

Lecture Recording for the purpose of remote learning began with a few studies where lectures were recorded and then viewings of the video footage were offered at specific times [1]–[3]. A subsequent approach was suggested by Goodhue et al. [4] where voice recordings accompany the slide sets used in a zoological lecture.

Major development in the field of ALR began when approaches to address the task were rapidly improved and iterated upon.

Keegan et al. [8] proposed the use of satellite broadcasting in a virtual classroom which could form part of open universities. Bodendorf et al. [9] also talk about broadcasting lectures as a "virtual classroom" for the purpose of remote learning so students are spared the time-consuming commutes between campuses. Soon after this, Chen et al. [12] discussed the potential of using low-resolution video streaming along with the lecture slides in HTML format to facilitate remote learning in the form of Web-Based Lecture technology (WBLT) and Mukhopadhyay et al. [13] explored the automatic creation of structured multimedia documents containing text, images, audio and video clips extracted from the live lecture. While these implementations are revolutionary and useful, they did not prioritise lecture recording in their research and so they do not offer much information about it.

The concept of Virtual Videography was introduced by Gleicher et al. [15] who identified the potential of recorded videos making use of videography and cinematography to enhance the viewing quality of automatically recorded lecture videos. The use of Videography and cinematography in recording lecture videos helps to make the recordings more visually pleasing to watch and so would be more readily accepted by students. It also makes the implementation more practical (e.g. ensuring the presenter is always in view, allowing the camera to focus on the most important parts of the video, capturing facial expressions or gestures, etc.) This idea was enhanced and iterated upon by many research endeavours since then and the pursuit of ALR began in earnest.

Liu et al. [16] took the advice from Gleicher et al. [15] and created an Automated Camera Management System which relies on cinematography guidelines to make its recording decisions. They interviewed professional camera operators in order to understand what should be followed when recording videos. Since lectures are not meant to be presented the same way movies are, the professionals gave them simplified guidelines for recording lectures. The researchers then created a program which would make use of these guidelines while deciding which camera to use in the venue at specific moments. They named this program the Virtual Video Director (VVD). The team found no significant difference between their VVD and the recordings made by the professionals when they conducted their evaluation but there was a slight bias for the recordings made by the professionals.

Rui et al. [17] described the development of a Virtual Cinematographer (VC) and a Virtual Director (VD) in a similar research endeavour. Their system used multiple VCs (a VC uses the professional camera operators' guidelines to make important camera-based decisions) and a single VD. The VD would use very high-level guidelines to decide which VC to use for each moment and the VC would know what to do given its context. Their findings were similar to that of Liu et al. [16] where users could not distinguish the program's output from the output created by professionals in a significant way. They acknowledged that there is more work to be done in this field and that there is potential in the use of ALR Systems.

Gleicher et al. [19] described a way to capture and edit lecture content using various cinematic principles and guidelines when recording lectures. They also surveyed the available computer vision tools available and then combined these tools with the guidelines in order to create a framework for a Virtual Videography system. They found a lot of potential in this framework and highlighted where further development is needed and that a fully functioning system should be implemented and tested. They also stated that Virtual Videography may not be a solvable problem.

Despite this, Rui et al. [21] published further research into ALR Systems. Their goal was to automate lecture broadcasts to remote audiences.

The 2 interrelated components of their system's design were the hardware/software and the rules, guidelines and heuristics required to produce aesthetically pleasing videos. They allowed others to build on their discoveries by explaining the video-production rules and acknowledged further research is necessary for more complex rules. They concluded by saying that automation could make a large impact on learning techniques and that automation costs are decreasing.

Soon after this Yokoi et al. [25] proposed a method to generate dynamic video from high-resolution images using a high-definition camcorder. The frames in the video were cropped to create a close-up effect on the presenter and they made use of bilateral filtering to smooth the transition between frames. They conducted a subjective experiment in which they found their algorithm to be effective. They showed the output of their program to a group of 20 students and each student was asked the same 5 questions about the video. These questions were as follows:

1. Is the ROI such as the instructor easy to see?
2. Is the motion by the instructor visible clearly?
3. Are the areas of the chalkboard that you are interested in easy to see?
4. Is the camera motion (panning) natural?
5. Is the lecture video comfortable to watch?

They found a significant difference in favour of their program between what they produced compared with a linear interpolation algorithm. While this research does not mention the use of cinematographic guidelines, it is useful because its implementation of an ALR system is indicative of the potential held by these systems in general.

Similar research by Zhang et al. [27] involved using a Pan Tilt Zoom (PTZ) Camera to record the coarse motion of a lecture and then, using video cropping, performed digital movements for smoother and finer control of the footage. They deployed a system called "iCam" which makes use of the program they developed. They consider their work to be a hybrid solution because it combines mechanical and digital camera motion. This research differs from ours since they make use of both mechanical and digital camera movements while we focus on just the digital movements. They also do their edits in real-time while our edits are done during post-production.

Heck et al. [29] built onto what was written about a framework for Virtual Videography by Gleicher et al. [19]. This research was comprehensive on the topic of ALR and it provides a lot of the heuristics a VC/VV system would use. There are 4 phases of VV mentioned in this paper:

1. Media Analysis - Finding region objects
2. Attention Model - Suggests areas of importance for framing
3. Computational Cinematography - identifies the most appropriate shot
4. Image Synthesis - Produces the output video

They mentioned that their system only works in a traditional lecture room environment but they hoped to expand the scope in future work. They covered the benefits of offline processing as well which are:

1. Better analysis
2. Better decision making
3. Better image synthesis
4. Able to produce different videos from the same source

They found that simple syntactic cues were generally enough for making framing decisions and stated that image synthesis is a useful alternative to specialised cameras.

Brooks et al. [39] conducted a study at the University of Saskatchewan to understand the popularity of WBLT. They created an online lecture presentation system and encouraged students to make use of it. They followed the students' progress on this system in order to see the impact that WBLT had on them and then asked them to complete a survey at the end of the 15-week term in order to gauge their perceptions of the system and their motivations for (or against) using it.
They found that, for traditional higher education approaches, the system audits and intervenes in student learning by allowing the use of analytics. They mentioned that intervention tools would be the subject of future work, however, they did various analyses of the data they gathered and found that the reasons students used the system were numerous and varying. They stated that the system produces interesting results even from the surface-level analysis. This research helps to indicate which students like WBLT and for which reasons.

Nagai et al. worked on the implementation of a high-definition lecture recording system for daily use [45]. They aimed to produce an ALR System, which would operate in post-production, cost-effectively. They found that long lead times and the lack of support for live streaming were problematic and that their virtual camerawork algorithm needs improving. The projector screens were sometimes too far to read clearly and their heuristics were not clear enough to choose the correct focus. They managed to reduce the cost of the system by combining a stationary High Definition camera with virtual camerawork and a file capture approach. They mentioned mobile computing technology as the target for future work. While the research team considers themselves unsuccessful in terms of the performance of their implementation, the cost-reduction is progress worth mentioning.

Wulff et al. [48] introduced an open-source system called "LectureSight" with the aim of making a free solution featuring active, automatic camera control in lecture and presentation environments. They conducted a study in which students from two universities studied for a synthesised exam using an ordinary lecture recording and another made using LectureSight. While it performed well in the study, they found there is still room for improvement and a psychological study indicates that lecture videos with a camera that follows the presenter are more beneficial since the presenter's gestures and facial expressions are more easily perceived.

A paper in 2017 [55] investigated the temporal aspects of studies designed to measure the performance of students using online learning resources. The study used seven online lectures posted on YouTube and the resulting analytic data for a period of four semesters. At the end of the study, it was observed that view counts only increased (on average) a week before the exams. This increase was observed for the view frequency and view duration and the increase in view counts occurred before each exam. This finding indicates that the online videos were considered as a meaningful and helpful resource by students.

Furini et al. [61] created a system called "ONELab", an online hosting platform built by the authors, to cater for students' online and remote learning needs. ONElab was deployed in five degree programs at the University and the enrolled students' performance data was recorded to measure its effectiveness. The results showed that ONELab users (on average) acquired more credits and scored better grades than non-users of the system.

2.4 Related Work

In some areas of an ALR System's design, there are many solutions available. In the following sections, we identify those areas and compare our work with what others have done in similar and relevant works. We focus our comparison to the VC module of the ALR System.

2.4.1 Real-Time Processing

While we haven't used a real-time solution, it is worth mentioning the merits of such an approach before moving on to the post-production approaches. Some note-worthy implementations of real-time solutions can be found in the works of Cavallaro, Chandrasekera, and Taj [28], Zhang, Rui, Crawford, et al. [34], Chou, Wang, Fuh, et al. [38], and Winkler, Höver, Hadjakos, et al. [42].

Real-time solutions run concurrently with the video recording process [38], [43] which means that there is no processing time required before the final product is ready (except the time taken to record the footage). A real-time implementation is usually paired with one or more PTZ cameras and the software instructs the camera how to move [28], [34].

Real-time solutions can only use information from past frames because future footage has not been recorded yet. Another restriction is that a real-time solution cannot use all the past information because the processing time must be minimised in order to meet the strict frame rate required for real-time processing [10], [24], [41]. Consequently, the confidence in decisions made by real-time solutions is potentially weaker than a post-processing approach. This means the proposed, post-processing VC can make decisions based on a much bigger context than a real-time solution would allow.

2.4.2 Post-Processing

There are many advantages to using a post-production workflow in an ALR System.

A post-processing implementation gives the VC access to the full video at all times [15]. This allows the VC to look at past and future frame information, which expands the context when making a decision [29].

Passing over the data multiple times allows the VC to separate its decision-making process into easier-to-understand steps and also helps to structure the way it applies its heuristics (e.g. making a full pass to process the presenter's locations and a (separate) full pass to process the board locations).

2.4.3 Camera Setup

There are many ways to set up the video capture system. The number and type of cameras used are variables which play an important part in setting the foundation of an ALR System (especially for the VC).

Camera types used in an ALR System generally fall into one of two groups, namely PTZ or Static. A PTZ camera provides mechanical motion and usually pairs with a real-time implementation [26]. Systems which use PTZ cameras to record their videos write software which is then used to drive the camera on-site and usually with an additional module [26], [43], [49]. A static camera does not have mechanical motion and, therefore, cannot move itself. Any shot changes must be made in post-production or by using multiple cameras between which a VC would switch [38], [43].

The number of cameras can change the way a VC operates significantly. While a single PTZ camera could cover the whole lecture venue, a system could use more than one to cover peripheral regions of a venue while multiple static cameras would remove camera motion completely [20], [42].

We decided on a single static camera with a wide-angled view such that the entire lecturing area is covered by the camera. This allows the VC to post-process the footage to look like a PTZ camera or human camera operator has recorded the video. Multiple Static cameras could also be used, but a single 4K resolution static camera is cheaper to implement in every classroom on campus and a single 4K static camera is also cheaper than a single PTZ camera [17], [30], [38]. This issue of camera cost is important and another reason why we chose a post-processing solution.

2.5 VC Heuristics

Regardless of the camera configuration used, a VC needs a set of heuristics to guide its decision-making. While machine learning algorithms could alleviate the need for heuristics [35], they require a lot of training and time to calibrate in order to get the desired output [59].

As a result of this, we chose to use heuristics and traditional algorithms instead of training a machine learning solution.

A VC should guide the viewer's attention to the most important information in a given context [15], [19], [29]. In a single static camera system, this means zooming in to include only the relevant content for the context. An important part of guiding the viewer's attention is proper framing [21], [29], [31], [57]. The VC should allow viewers to see the presenter's face and gestures (where relevant) and any boards which apply to the context.

If a presenter moves a lot, the VC should zoom out or switch to a camera with a wider view of the venue to avoid jittery motion or frequent transitions so as not to disorient or confuse viewers [7], [15]. In cases where there is no board and the presenter is not moving a lot, the VC should tighten its view to a closer shot (called a close-up in the book by Katz [6]). If the VC loses the presenter, or the presenter is not in the camera's field of view, it should zoom out such that the viewer does not notice a problem (which would translate to a medium or full shot in the book by [6] depending on the context).

In Rui, Gupta, Grudin, et al. [23] there is a series of rules used for lecturer tracking and camera framing. Many of the aforementioned heuristics are also mentioned here while adding that there should be sufficient space above the presenter's head (described as a 10 - 15 cm space). The authors also mentioned that the lecturer should be centred in the shot whenever possible but (if the context requires it) allow lead room for the presenter's motion or use of the boards in the scene. The number of shot changes or camera movements must be minimised wherever possible [7], [15], [19], [22] as this prevents visually disorienting or confusing footage. Any camera motion should accelerate to constant speed and then decelerate to a stand-still in a "slow in, slow out" style for each shot transition [5], [11] which stabilises the motion by reducing jittery and disorienting movements.

We decided to use all the aforementioned heuristics in our implementation of the VC with the exception of the following:

We were unable to use gesture information (despite its usefulness for framing composition) due to the limitations of temporal differencing as a reliable means of identifying gestures from the presenter's motion. The tracking front-end used temporal differencing as its chief technique for motion detection and it was unable to track gestures reliably with this technique.

We also did not follow the rule where the presenter should remain in the centre of the shot wherever possible since this would mean the VC would make more transitions than it would if it limited motion only to when the presenter comes close to the boundaries of the current framing choice. This change ensures that the VC can keep the shot stable enough to allow viewers to follow what is being written on the boards.

2.6 Chapter Summary

This chapter begins by providing an overview of the background of ALR while going into detail on the various aspects of the field, as well as covering the history of the ALR and Virtual Cinematography fields.

We also discussed related research and how their work has influenced our VC in the following ways:

We chose a post-processing system, giving the VC access to the whole lecture to make decisions based on past and future events rather than only using past information to predict future action. This allows more time for framing choices with higher confidence levels and the VC can take multiple passes over the data to gather information on the various contexts in the lecture.
Our system is designed to run on a static camera with a wide-angle view of the lecture venue. This allows the VC to crop the most relevant information per frame in such a way that the resultant video appears to have been recorded by a human camera operator in the first place.

We used the related works in this chapter to inform our choosing the VC's heuristics.

Chapter 3

The Virtual Cinematographer Framework

This section discusses the design of the VC module and describes the module in terms of the data the VC expects as input, the processes the VC executes on the input data and the format of the output produced by the VC.

Figure 3.1 illustrates how the VC module fits into the system as a whole and more detailed figures are included below that illustrate the internal processes of the VC module in more detail. The system comprises modules that communicate using the pipe-and-filter design pattern (Ohlemacher [52]). By using a pipe-and-filter system with decoupled modules communicating only via a structured data file (JSON) TRACK4K is able to run each module on a separate machine (potentially) with dedicated resources. It also means multiple videos could be running concurrently at different stages in the pipeline (depending on how many machines an institution has allocated to processing lecture recordings). The modularity of the system makes integration easy.

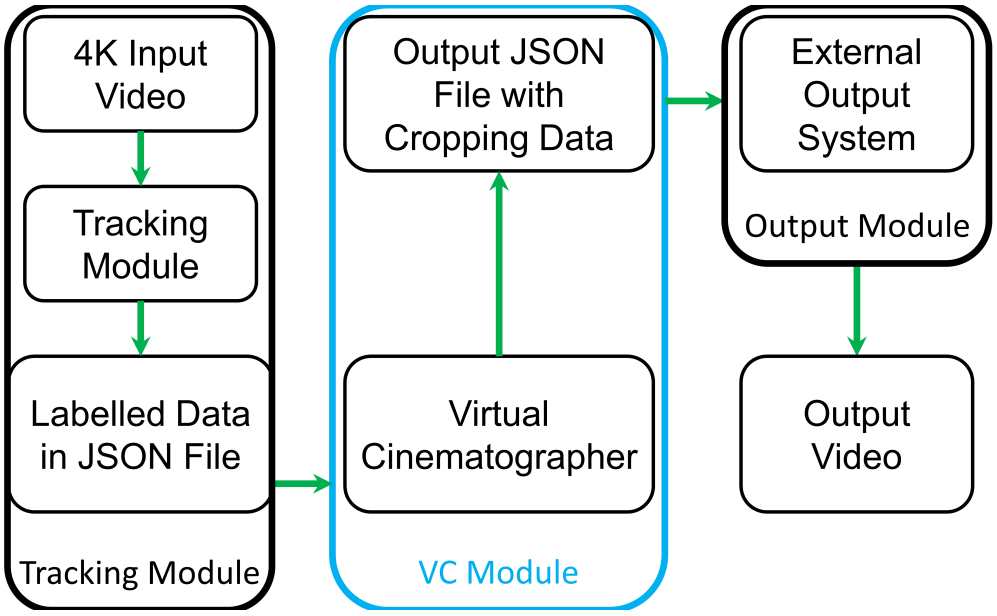


FIGURE 3.1: High-level system diagram of the TRACK4K program

The VC takes in the output of the Tracking Module (a separate module from this project), modifies it according to its internal processes (discussed in more detail below), and then sends the resultant data as a JSON file to third-party software for video production.

Decoupling the modules in this way means the output JSON file could be used on the client side. It also means that the VC is able to manipulate the input data in a non-destructive way because the original information is stored in an external file which is only accessed for reading.

A decoupled system is also easier to integrate during development, allows the modules to be added to other systems more easily, and can be distributed over multiple virtual machines on a server.

For example, a client-side video player could use the cropping information provided by the output JSON file and crop the original footage as it is being played back to viewers without generating a new video file. Another example would be to delegate the generation of the output video to a separate virtual machine on the server before distributing it to the client-side video players.

3.1 VC system input

The VC takes labelled, structured data, from a JSON file, as input from the tracking module. The figure below is a visualisation of the information provided in the JSON file from the tracking module.

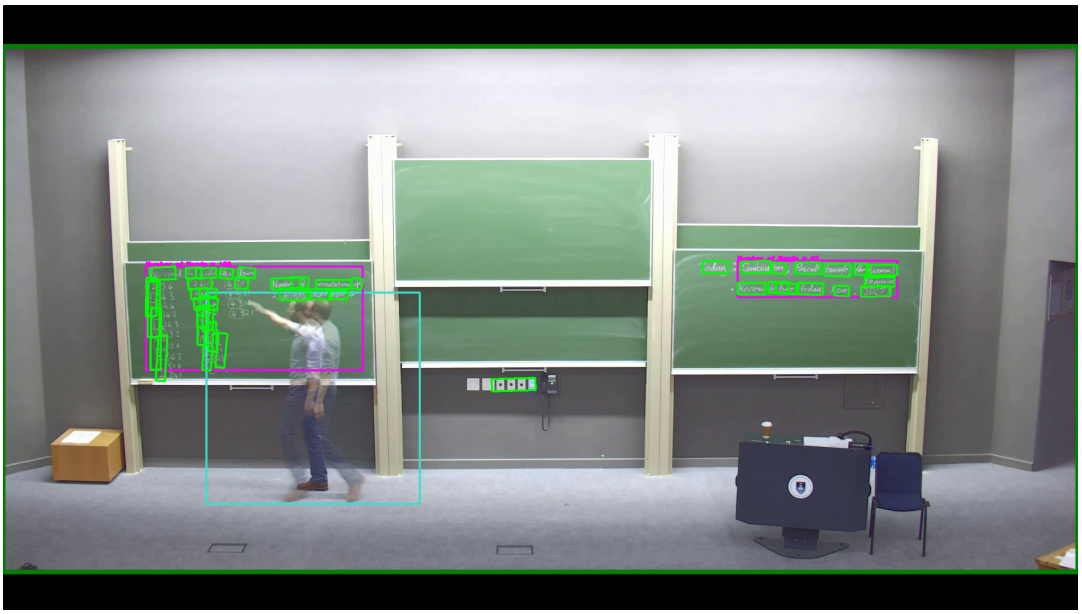


FIGURE 3.2: Visualisation of the labels attached to each frame by the Tracking Module

As shown in Figure 3.2, the data includes the coordinates of the presenter per video frame, stored as the top-left and bottom-right vertices of the presenter's bounding box. The tracking information also lists the bounding boxes of the boards in the venue (the purple boxes), the number of words on those boards (represented by the green boxes) and a feature count is also listed in the JSON file for each board in a frame. The presenter's gesture information is included (if available) by indicating the direction of the gesture and providing its bounding box.

The black bands in this screenshot are privacy masks that are added to the videos to protect the identity of the students captured unwittingly in the recording area, or to remove regions of the camera view that never contain useful information for both the tracking module and the VC. The VC does not include the area covered by the privacy masks in its output, which means the working area is reduced by however much of the footage is masked away. By doing this, the maximum level to which the VC can zoom out is reduced because the video's aspect ratio must remain the same throughout the video to avoid causing visual distortions.

3.2 VC system processes

As the input file is being read in, the VC organises the information into its internal representations for later use. The VC is divided into separate classes to process the input data more manageably. There is a class that stores the information for the presenters and another to store the information for the boards.

The VC takes in a JSON file from the tracking module and makes no further communication with it throughout the process. We will now discuss the data structures, classes and methods used by the VC to complete its tasks in the system for a lecture video.

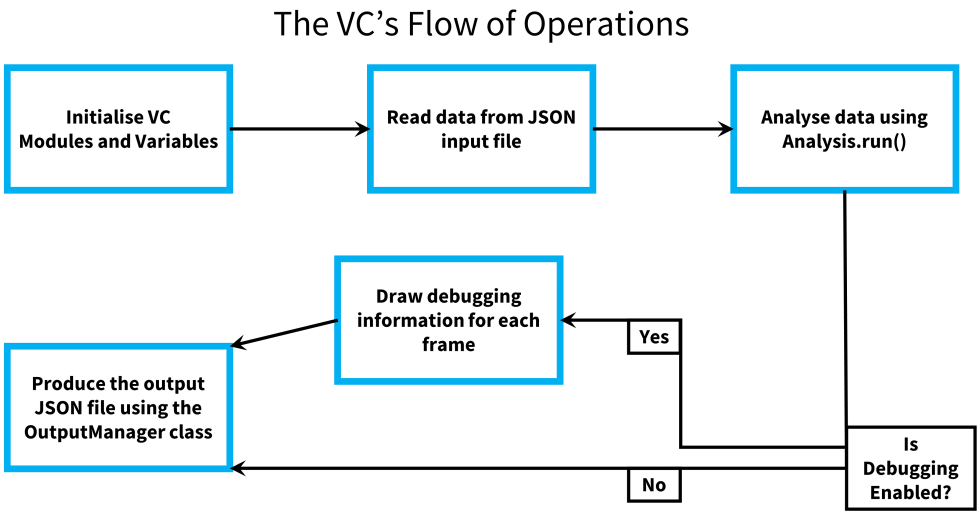


FIGURE 3.3: Flow diagram showing the high-level flow of operations

Figure 3.3 shows the flow of operations as the VC runs through a video. The VC begins by initialising its methods, modules, and variables. Once the VC is ready, it reads the JSON file from the Tracking Module and saves the information into its newly initialised data structures. The next step is to analyse the newly stored data such that the VC has a context in which to make decisions on how best to frame the presenter. If the flag for debugging is enabled, the VC adds debugging information to each frame before sending it to the OutputManager class. The OutputManager class writes the output file containing the cropping information per frame such that third-party software, such as FFMPEG, can create the final output video.

We discuss this process in more detail in the following sections.

3.2.1 Serialise data from the JSON file into an internal representation

This part of the process is short and simple. The VC expects data from the tracking module to populate its classes. It knows which variables in the input JSON file correspond to each of its internal representative data structures. During this process, the VC iterates over the information in the JSON file and saves the information to its internal representation, which it stores in memory for quicker access. The VC then closes the JSON file and does not need to access its data again.

The following information is extracted from the start of the JSON file:

- The name of the input video file
- The frame rate of the input video
- The total number of frames in the input video
- The maximum width and height of the input video

- The interval of frames sampled by the tracking module
- The resolution with which the tracking module processed the data (called co-ordinate space)

Once this initial information has been processed, the VC starts reading the data for each frame and storing it in its internal variables, and data structures.

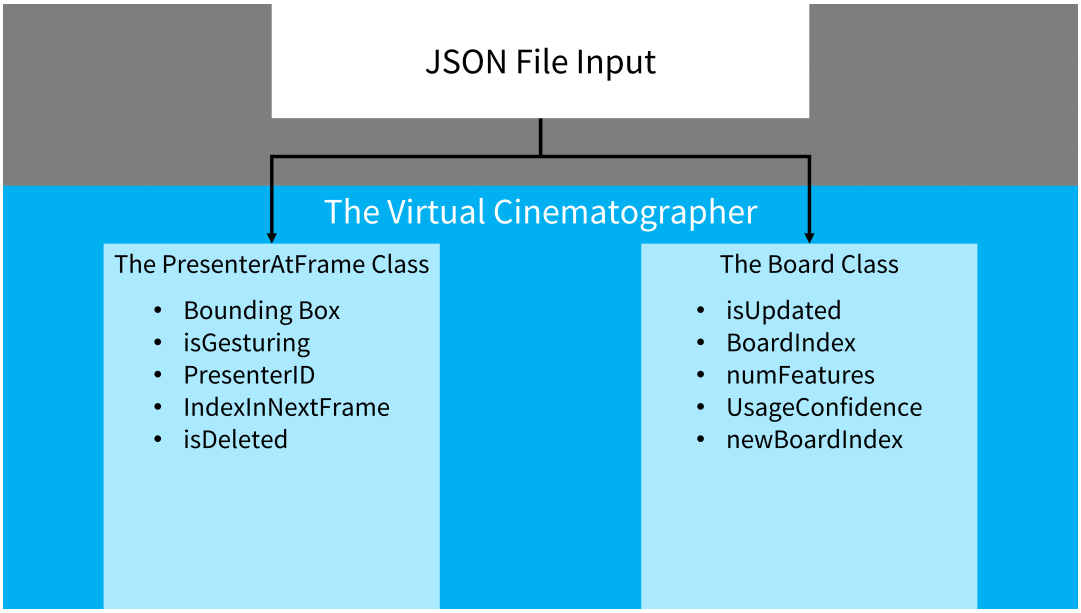


FIGURE 3.4: The VC splits the JSON input into the PresenterAtFrame and Board Classes

After reading all the tracking information, the VC restructures the data into a format that makes it easier to process internally. All the information about the presenter is consolidated into the *PresenterAtFrame* class so that the VC only has to process a collection of objects from this class. Data like the bounding box on each frame, where the presenter was labelled as being in view, and the gesture information per frame for that Presenter are included in this class. There are a number of constructor methods to help the creation of objects for this class under different conditions.

This is repeated for each board in the venue by creating a separate object for each board from the *Board* class. This class stores the bounding boxes for each frame on which the board appears and the number of features visible on the board for that frame.

3.2.2 Data analysis and decision making

The next step after restructuring and storing the input data is to run several processes on the data to understand the VC's environment as a context for making high-level decisions later.

Figure 3.5 illustrates the flow of operations during the Analysis.run() method. The first steps are to fill in frames with missing presenter information and to check for any board activity. With this information, the VC can then separate the information into discrete sections, called Contexts, and deal with each Context appropriately.

A Context is a sequence of frames grouped by presenter and board activity in such a way that the VC does not need to perform a transition from one frame to another in the sequence.

When addressing a context, the VC must decide if the camera should adjust its zoom level such that it includes more (or less) information based on what is relevant in that context. The last step is to create the transitions over each context by setting the cropping window for each frame in the context such that the resultant section of the video appears to have a camera transition across the lecture venue.

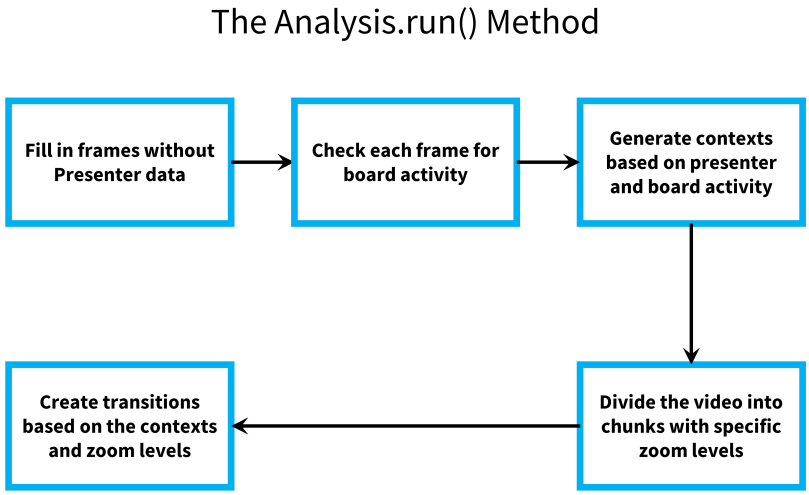


FIGURE 3.5: Flow diagram showing the sequence of operation in the Analysis.run() method

We discuss these steps in more detail in the sections below.

3.2.3 Interpolating the gaps in the data

Sometimes a sequence of frames from the lecture venue is marked as a busy scene by the tracking module (e.g. when there are multiple presenter candidates or the audience comes to the front of the venue during the lecture and the true presenter cannot be resolved). The data about the presenter is not stored for frames in busy scenes.

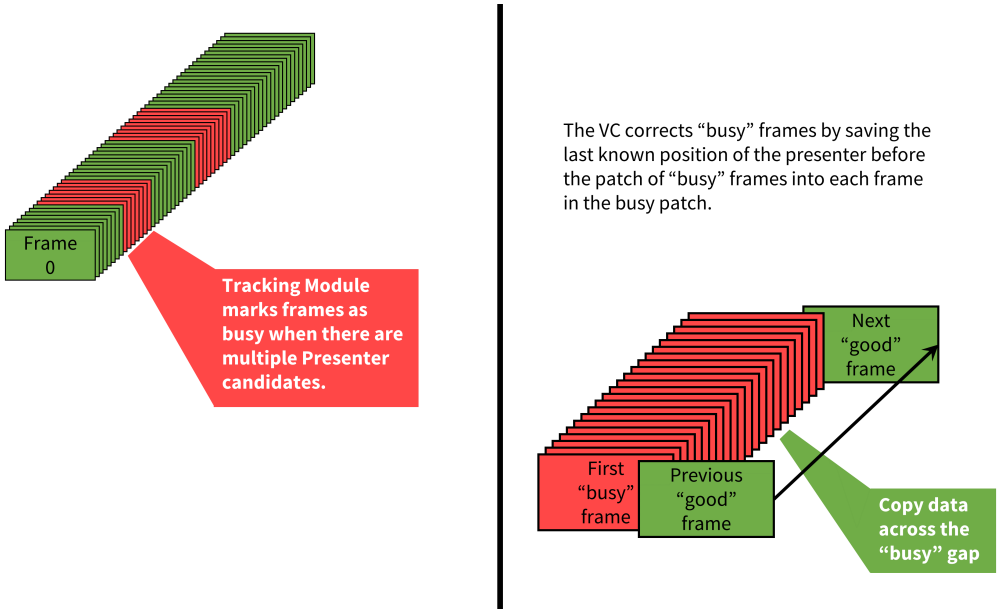


FIGURE 3.6: The VC must fill in the gaps in the tracking data

The VC must, therefore, find a way to work around these gaps in the data to make better high-level decisions about framing the content according to the context. Later steps in the process use the presenter's location (and bounding box) to compare with other, simultaneous changes in the venue for better framing (e.g. how the features change on a nearby board).

The solution is to populate the frames in the gaps with the last known position of the presenter. The VC will create transitions in later stages of the process to interpolate these discrete changes in position.

Iterate over video to ensure all frames where presenter is expected contain presenter data.

When such a frame is found, save the previous frame's data into it to fill the gap.

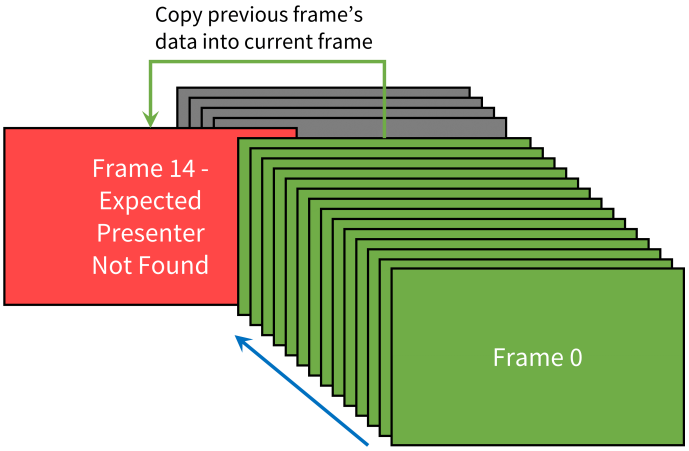


FIGURE 3.7: The VC iterates over frames without presenter data and fills in the last known position for those frames

```
1.  FOR EACH ContextFrame IN contextFrames:
2.      FOR EACH Presenter IN presenters:
3.       IF NOT Presenter.deleted:
4.           accumulate_mean_width()
5.           accumulate_mean_height()
6.  FOR EACH ContextFrame IN contextFrames:
7.      IF ContextFrame.presenters.empty() OR ContextFrame.presenters.deleted():
8.       IF NOT ContextFrame.has_previous():
9.           presenters.createFakePresenter(videoCentre, meanWidth, meanHeight)
10.      ELSE:
11.          contextFrames.copy_last_frame()
```

FIGURE 3.8: The algorithm of iterating through the presenter data and filling in the gaps

The VC starts by iterating over each frame to ensure there are no frames with missing presenter information. In some cases, the VC expects presenter information from the tracking data but the tracking data for the presenter is missing for a group of frames. The VC expects the presenter's information based on contextual clues such as the presenter's trajectory across the venue not slowing down or stopping before the presenter's information is lost or if the presenter's data is only missing for a short interval where a presenter could not have left the venue and returned realistically. The last known position of the presenter and the next known position of the presenter also contribute to the trajectory.

After these gaps in the presenter's tracking information are identified, the VC fills in the last known position of the presenter for each frame in the gap until the VC encounters a frame where the presenter's location is, once again, available. If the first frame has no presenter data, the VC creates a false presenter in the middle of the venue and copies that frame into subsequent frames with missing presenter information.

3.2.4 Determine board usage

The VC also checks the board activity at this stage in the process. It does this by checking each frame in the video to see if it contains any board information.

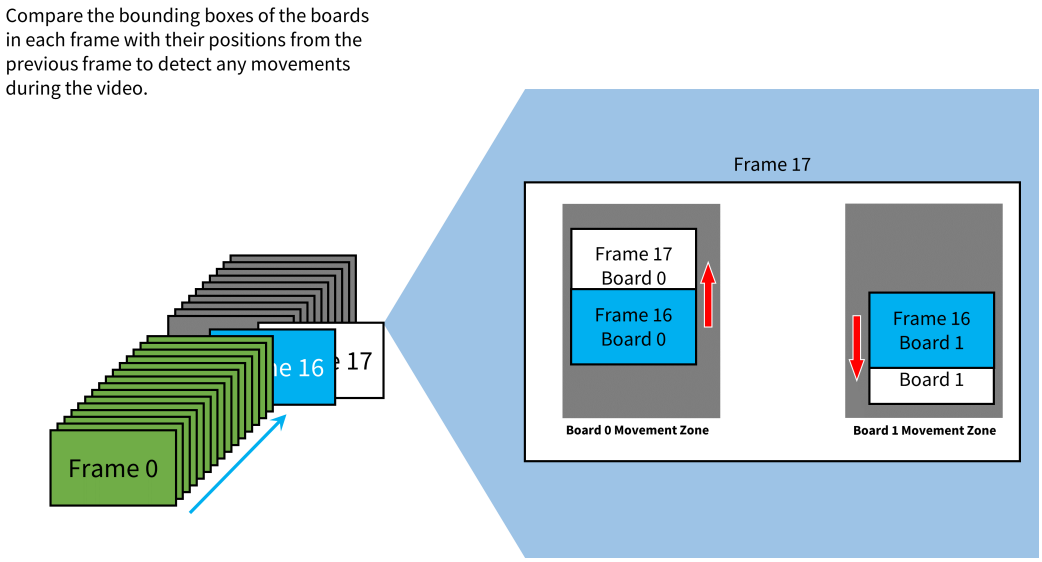


FIGURE 3.9: The VC compares the boards of the current frame against the boards in the previous frame

If there is board information on that frame, the VC iterates over all the boards and checks which ones overlap with boards from the previous frame. This is because the VC assumes boards in lecture venues can move, so the VC must make sure the boards labelled in the current frame are the same boards that were labelled in the previous frame. The tracking module saves boards based on the features it finds on them. This means that the recorded board can be a lot smaller than the actual board in the venue. In some cases, the tracking module will record multiple boxes for a single actual board in the venue. In these cases, the VC treats each box as a separate board. This comparison also helps the VC to identify cases where the presenter is standing in front of the boards because the tracking module will change the size of the bounding boxes of boards in such cases. This continuity is then used to determine board usage over the length of the footage.

Figure 3.10 shows the flow of operations performed by the VC to determine board usage in a lecture video after the VC performs the board overlap comparison. This process is performed on each board in the venue sequentially.

Overall description

There are 5 steps in this process (as shown in the figure above): Step 1 determines whether the tracking data includes any boards. If there are boards, the VC must separate the video into board usage contexts (Step 2). Step 3 combines video segments with a consistent change in the number of features per board (described in more detail below). These video segments are known as board usage contexts and are used in Step 4 to determine board usage by following trends related to the feature counts on the boards. Boards with a large change in the feature count over a board usage context are more likely to have been modified than boards with a smaller change in feature count.

Step 5 combines the board usage with other information at the same time interval in the video to determine which of the modified boards are relevant to the overall context of the video. Each step in this process is covered in more detail below.

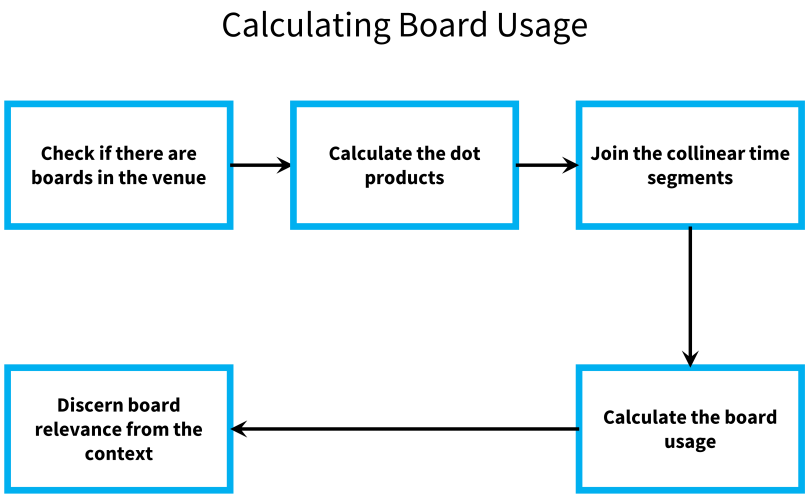


FIGURE 3.10: The flow of operations by the VC to determine board usage

Detailed description

We plot all the feature counts of a board on a graph where time is represented by the frames in the footage. We must separate the footage into Contexts (as defined above) which require the VC to respond differently from neighbouring Contexts. We begin this segmentation by dividing the footage into time intervals of the same, small number, of frames. This frame interval must be set up such that it is a whole number for which the total number of frames is a multiple to prevent intervals from being shorter or longer than others. For each interval, we plot a line of best fit to represent the change in the feature count for that section of time. Once we have plotted all these lines, we compare the gradients of neighbouring lines from left to right using the dot product. By comparing gradients we can detect changes in the number of features.

If Line A $= (a_1)$ x $+ (a_2)$ y and Line B $= (b_1)$ x $+ (b_2)$ y
Then A.B $= (a_1)(b_1) + (a_2)(b_2)$ or A.B $= |A||B|\ Cos\theta$

The closer the result is to 1 (or -1) for the dot product between two neighbouring lines, the more alike they are in their gradients, which means these lines are very likely part of the same context (which we will call approximately *collinear*). The closer these lines are to 0, the more likely they are to be perpendicular and it is more likely that these lines are from different contexts. To reduce noise and increase confidence in the approximate *collinearity* between neighbouring line segments, we use a threshold such that values of $x >= 0.9$ are classified as approximately *collinear* and all other values are classified as distinct line segments. We used a *collinearity* score of $x >= 0.9$ as a threshold to reduce noisy board usage information. By reducing noise in this way, the VC can have high confidence in the board usage contexts and the decisions it makes from those contexts.

Collinear intervals are then combined into larger time intervals of the video and stored as a single context. We include more information about the dot product and how we use it to determine *collinearity* between line segments in section A.1 of Appendix A.

Figure 3.11 shows the algorithm used by the VC to calculate the dot products for each video interval. The VC saves the frame number and the feature count of the board for every frame in the current video interval. Once the feature counts are saved, a line of best fit is drawn for each video interval and saved to a vector. The VC then calculates the dot product between neighbouring lines of best fit and stores those in a new vector.

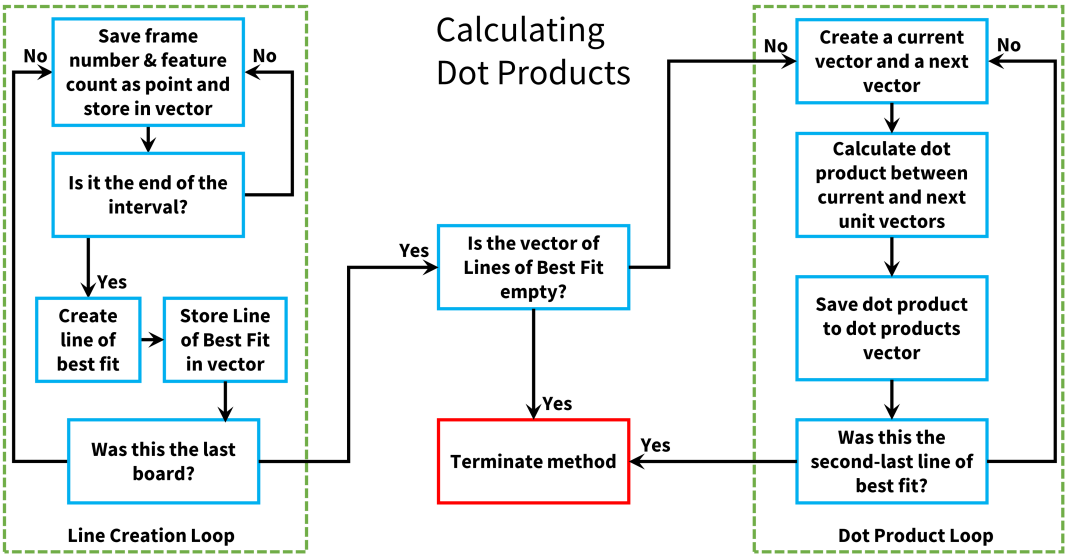


FIGURE 3.11: The algorithm used to calculate the dot products for each video interval

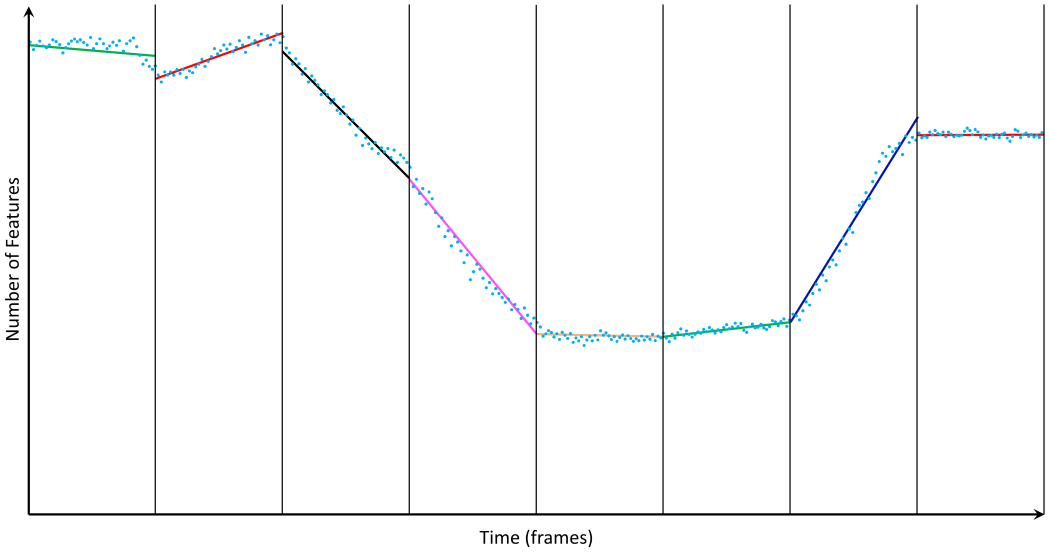


FIGURE 3.12: The visualisation of the lines of best fit in the 'calculateDotProducts' algorithm

Figure 3.12 is a visualisation of the lines of best fit mentioned in the algorithm above. The vertical axis represents the number of features in a frame, while the horizontal axis represents the frames in the video. The columns in the graph represent the video intervals mentioned above.

A new line of best fit is created for each of the intervals such that a gradient may be obtained from that line. In intervals 3, and 4 in the figure above, we see that the black and magenta lines are very close to parallel, so they should be combined. Having merged intervals 3 and 4, the VC will detect the number of features on the board is decreasing rapidly, which means the board should be marked as relevant. Since we assume the lecture venue is quiet, and the presenter is the only one moving around during the lesson, the VC will assume the presenter is the one changing the feature count on the board (either by writing new content, occluding the boards, or erasing the existing content).

In Figure 3.13, we see that neighbouring line segments have been combined into larger line segments when they were approximately *collinear*, which represent a new, larger context. The VC requires more information to determine what caused the change in the feature count since the number of features per frame does not provide enough information.
This stage in the process merely groups frames into events which will be combined with other aspects of the scene to determine the full context.

If we use Figure 3.13 as an example, the VC might see the sharp drop in the feature count during the grey-coloured section, followed by the stable orange-coloured section, and combine this information with the presenter's location to conclude that the presenter has either erased the content on the boards or is occluding it. If the VC looks even further ahead and takes both the blue-and-red-coloured sections into account as well, there is more evidence to indicate the content was merely being occluded, since the features returned later in the video (unless the content is being displayed on a projector screen).

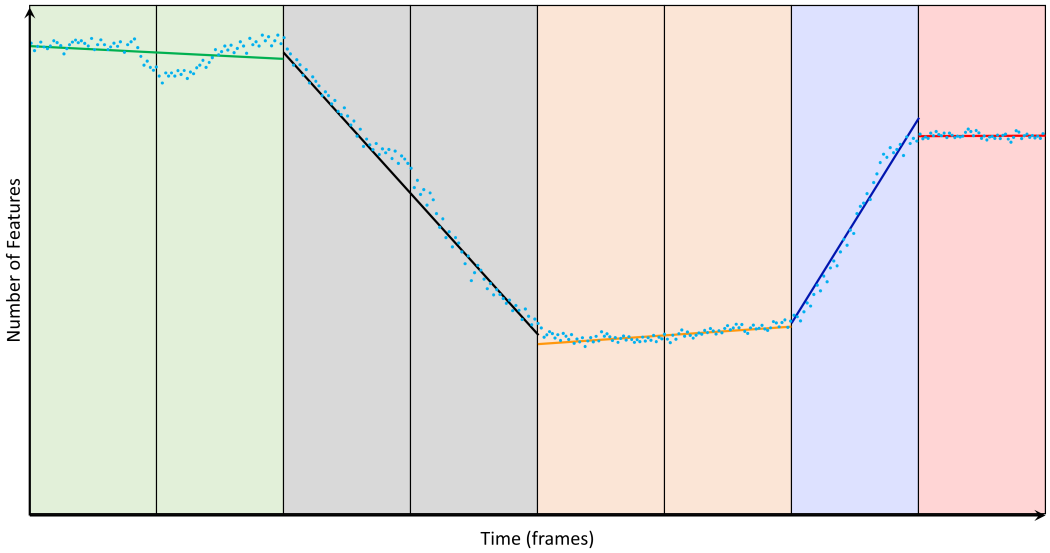


FIGURE 3.13: The visualisation of the lines of best fit in the 'calculateDotProducts' algorithm

```
1.  IF NOT dotProductVector.empty():
2.      Vector<VideoInterval> tempDotProductVector
3.      FOR i IN dotProductVector:
4.          VideoInterval VidInt(dotProductVector[i].START, dotProductVector[i].END)
5.          IF dotProductVector[i] > 0.9 AND dotProductVector[i+1] > 0.9:
6.              WHILE dotProductVector[i] > 0.9 AND dorProductVector[i+1] < 0.9:
7.                  IF i > (dotProductVector.size() – 1):
8.                      BREAK
9.                  VidInt.END = dotProductVector[i].END
10.                 i++
11.         tempDotProductVector.push_back(VidInt)
12.     dotProductVector = tempDotProductVector
```

FIGURE 3.14: The algorithm used to join collinear segments

The next step is to use the newly merged time intervals to determine board usage. The VC creates a new line of best fit from the feature counts of each frame in the new time intervals.

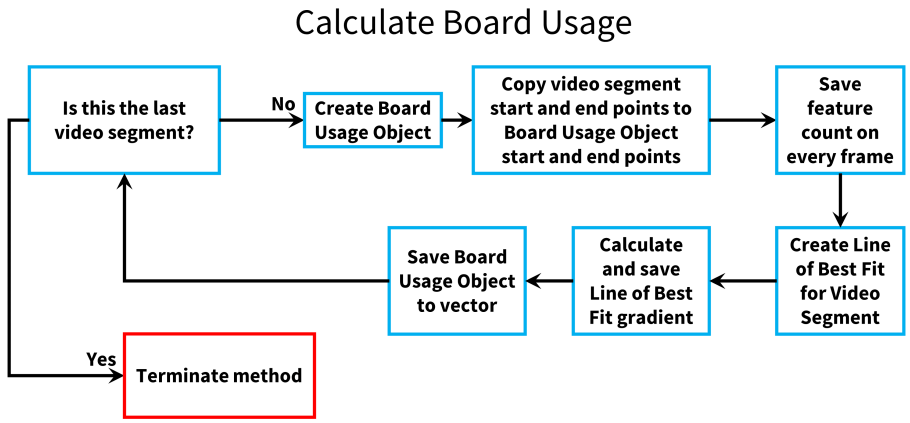


FIGURE 3.15: The flow of operations in the calculateBoardUsage method

Usage is determined by calculating the gradient of the line. A positive gradient, for example, means there was an increase in the number of features on the board over the video segment which means content is being added to the board. The VC must now identify what each gradient means in the overall context of the video by looking at more than just board data.

The last step in the board usage detection phase of the VC's analysis is to determine which boards, highlighted as "in use", are actually relevant to the overall context. This step will guide the VC to the most important area of the venue for each frame and help with framing decisions.

```
1.  FOR i IN range( boards.size() ):
2.      FOR EACH Usage in usages:
3.          FOR j IN range( Usage.startFrame, Usage.endFrame ):
4.              presenter = contextFrames[ j ].get_presenter()
5.              average_rect = boards[ i ].averageRectangle
6.              intersection = calculate_intersection(presenter.boundingBox, average_rect)
7.              midpoint = get_midpoint(presenter.boundingBox)
8.              leftHorizontalDistance = abs_value(presenter.boundingBox.top_left.x –
                presenter.boundingBox.bottom_right.x)
9.              rightHorizontalDistance = abs_value(average_rect.top_left.x – average_rect.bottom_right.x)
10.             maxHorizontalDistance = max(leftHorizontalDistance, rightHorizontalDistance)
11.             height = getAspectRatioCompliment(maxHorizontalDistance)
12.             IF intersection.area() > 0:
13.                 contextFrames[ j ].usageRelevant = TRUE
14.                 Usage.relevant = TRUE
15.             contextFrames[ j ].usageRelevant = height <= getGlobalHeight()
```

FIGURE 3.16: The first of the discernBoardRelevance() method

The pseudocode in Figure 3.16 is the first part of the discernBoardRelevance() method. The VC iterates over all the boards and then goes through each usage recorded per board. In order for a board usage to be marked as relevant, it must be close to the presenter at the time it was recorded as being "in use". This check is performed by testing the intersection of the presenter's bounding box with the average rectangle of the board (see Figure 3.17) over the whole of the video.

This average bounding box is calculated for each time interval and overwritten for each frame in the interval. The VC uses the average bounding box of the board since the size and position of the board change over the length of the video.

If the area of the intersection is greater than zero and the height of the intersection is shorter than the global height of the venue, then the usage is marked as relevant.

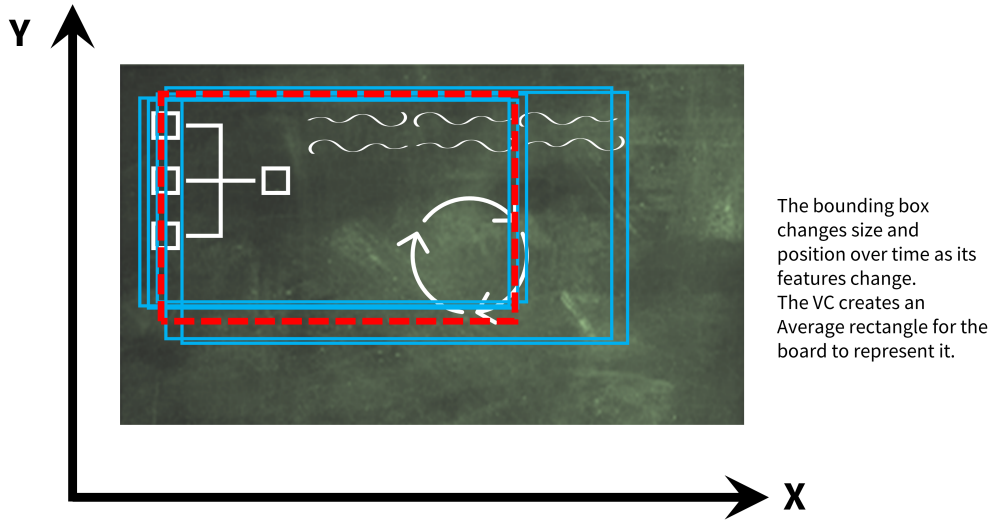


FIGURE 3.17: An average bounding box is calculated to represent the board for the time interval

```
1. FOR i IN range( boards.size() ):
2.     currentBoard = boards[ i ]
3.     vector<Usage> newUsages
4.     FOR EACH Usage IN boards[ i ].usages:
5.         FOR j IN range( Usage.startFrame, Usage.endFrame ):
6.             newStartingIndex = 0
7.             newEndIndex = 0
8.             IF contextFrames[ j ].usageRelevant:
9.                 newStartingIndex = j
10.                WHILE contextFrames[ j ].usageRelevant:
11.                    newEndIndex = j
12.                    j++
13.                Usage tempUsage
14.                tempUsage.startIndex = newStartingIndex
15.                tempUsage.endIndex = newEndIndex
16.                newUsages.push_back(tempUsage)
17.            j++
18.    boards[ i ].usages = newUsages
```

FIGURE 3.18: The second of the discernBoardRelevance() method

The pseudocode in Figure 3.18 is the second part of the discernBoardRelevance() method and it follows directly after the first part. It deals with modifying the board usage intervals for each board based on the relevance of those board usages. The VC iterates over the boards in the venue and checks each of their usages.

For each usage, the VC iterates through each frame recorded as part of its interval. The VC then adjusts the usage interval to include only the frames that were marked as relevant.

3.2.5 Generating video contexts

Now that the presenter is recorded for each frame in the video and the VC has separated the video into segments of relevant board usage, the VC can now organise the video into contexts using this information to make it easier to decide on how best to frame the presenter within that time interval.

```
1.FOR i IN range( contextFrames.size() ):
2.    topY = INT.getMax()
3.    topX = INT.getMax()
4.    bottomX = INT.getMin()
5.    bottomY = INT.getMin()
6.    FOR Presenter IN presenters:
7.        IF Presenter.deleted():
8.            IF Presenter.boundingBox.topLeft.x < topX:
9.                topX = Presenter.boundingBox.topLeft.x
10.           IF Presenter.boundingBox.topLeft.y < topY:
11.               topY = Presenter.boundingBox.topLeft.y
12.           IF Presenter.boundingBox.bottomRight.x < bottomX:
13.               bottomX = Presenter.boundingBox. bottomRight.x
14.           IF Presenter.boundingBox.bottomRight.x < bottomX:
15.               bottomY = Presenter.boundingBox.bottomRight.y
16.   contextFrames[ i ].framingRect = create_rectangle(TopX, TopY, BottomX, BottomY)
17.   contextFrames[ i ].framingRect.lift_enclosingContext(globalCropHeight * 0.2)
18.   FOR j IN range( boards.size() ):
19.       IF boards[ j ].usageConfidence > 0.0:
20.           grab_nearby_boards(contextFrames[ i ].framingRect)
```

FIGURE 3.19: The algorithm used to create contexts using the correct framing

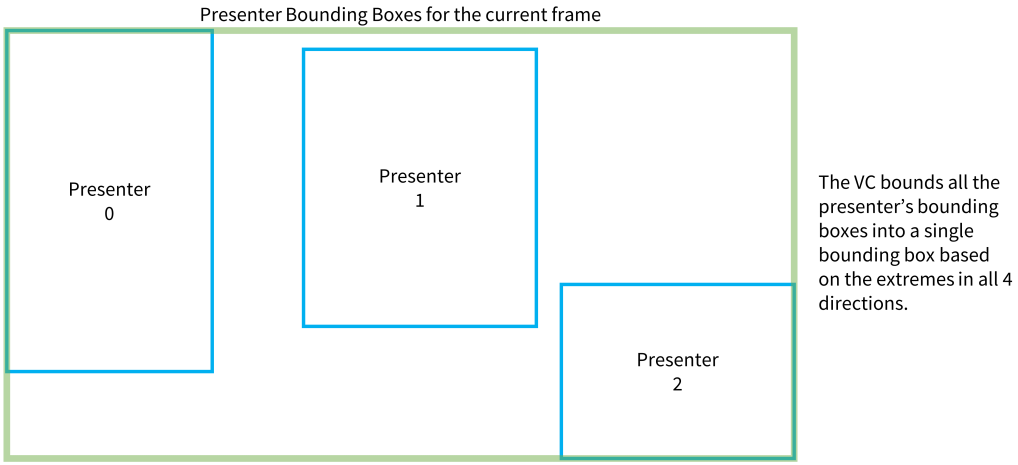


FIGURE 3.20: Potential presenters' bounding boxes are grouped together

The VC combines the bounding boxes of all the potential presenters per video frame to determine the position and scale of the cropping window.

This is done as the first step when there are multiple presenters recorded in the tracking data or if there is no confidence in the presenter information. The VC's priority is to frame the presenter, or as many presenters as possible in the case of multiple presenters (see lecturer tracking and framing rule 2.2 in [23]). The VC does this by recording the leftmost, rightmost, highest and lowest corners from all the presenter candidates' bounding boxes for the current frame. Once recorded, the VC creates a rectangle to represent the area within which all the presenters are enclosed. This rectangle is called the 'Enclosing Context' and it represents the focal area for the VC on that frame.

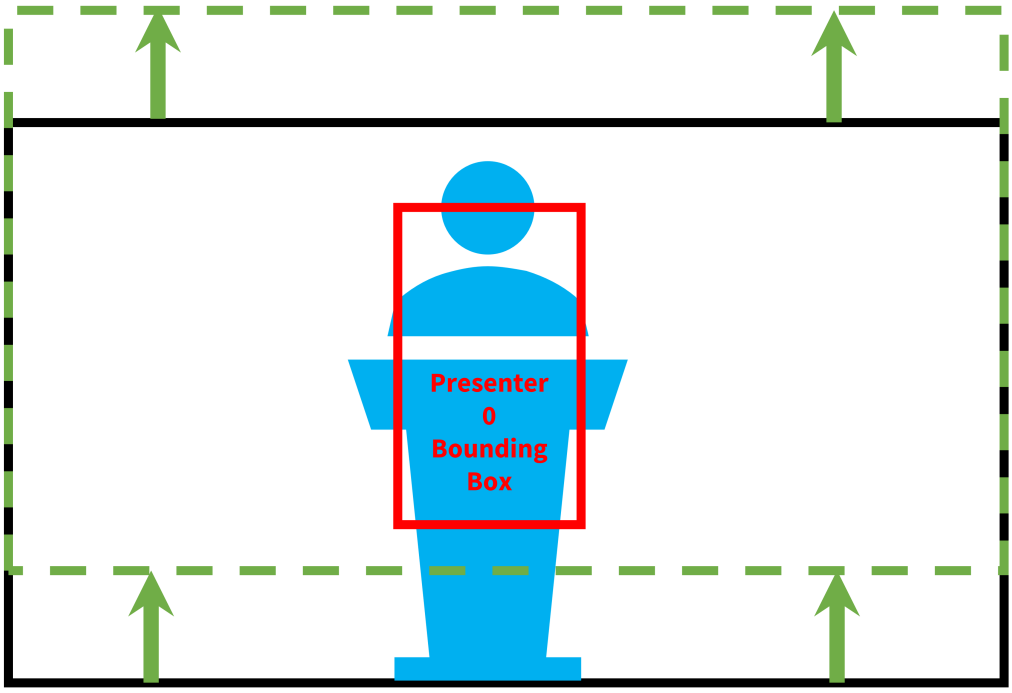


FIGURE 3.21: The cropping window is lifted slightly to allow more space above the presenter's head

As a matter of good framing practice, the VC also lifts this new enclosing rectangle slightly to have some space above the bounding box of the presenter (see lecturer tracking and framing rule 2.1 in [23]). Not only does this account for cases where the bounding box did not enclose the whole presenter, but it also leaves some space above the presenter for a close-up view that does not seem cramped to the viewer. In the example figure above, the bounding box of the presenter does not cover the whole of the presenter's body. There is also not a lot of space above the presenter's head, so the VC moves the enclosing context up to give the presenter more overhead room (despite the potential loss at the bottom).

The next step is to check if there are any boards near the presenters for each frame (as shown on lines 18 - 20 of Figure 3.19). The VC checks if the bounding boxes of the boards in use for the current frame overlap with any of the presenters' bounding boxes. If there are any boards that overlap, then the Enclosing Context is expanded to include the bounding box of these boards as well. This step fulfils the heuristic of guiding the viewer's attention to what is important as was highlighted in previous works [14], [19], [29].

Once the Enclosing Context is expanded to include the bounding boxes of the nearby boards in use, the VC must adjust the aspect ratio of the Enclosing Context to match the cropping window's aspect ratio and to keep the aspect ratio consistent throughout. The Enclosing Context is then set as the cropping window for the current frame.

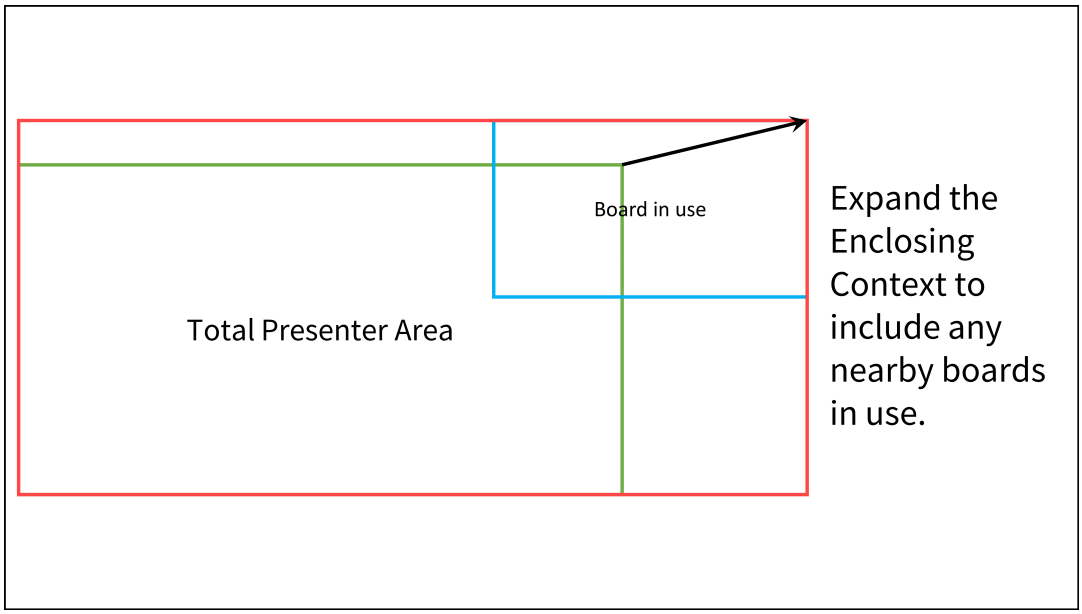


FIGURE 3.22: The enclosing context is expanded to include the nearby boards being used

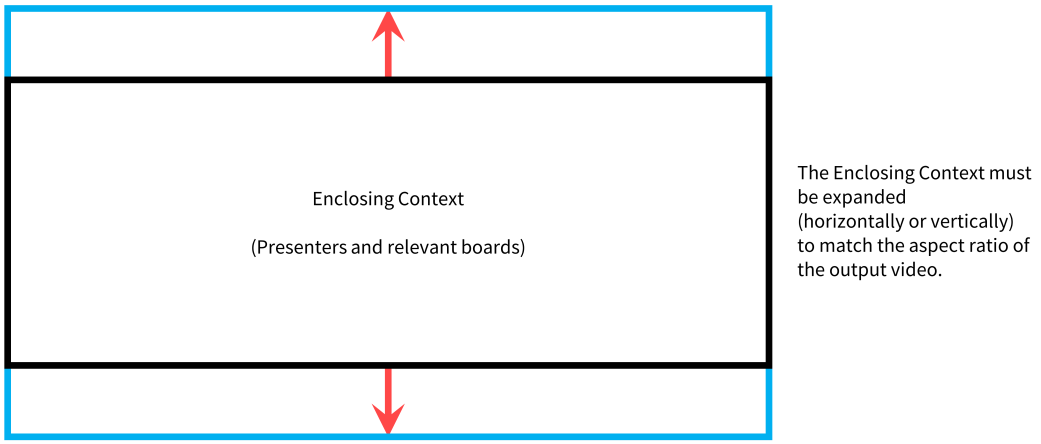


FIGURE 3.23: The enclosing context aspect ratio must be corrected after it is expanded

3.2.6 Generate zoom levels

Once the VC has established the Enclosing Contexts and adjusted the aspect ratio of the cropping window, it must determine how the VC will transition between these rectangles.

Figure 3.24 illustrates a case where the presenters' movements and nearby boards that are marked as relevant cause the Enclosing Contexts for each frame to differ in size and position over time. The cropping windows, at this stage in the process, would result in a video that is too jittery for viewers. The VC must therefore reduce the noise in the size and position of these cropping rectangles. To keep good framing practice, the VC separates the video into distinct zoom levels between which it will transition in later steps of the process.

It does this by grouping the enclosing contexts based on their height values.

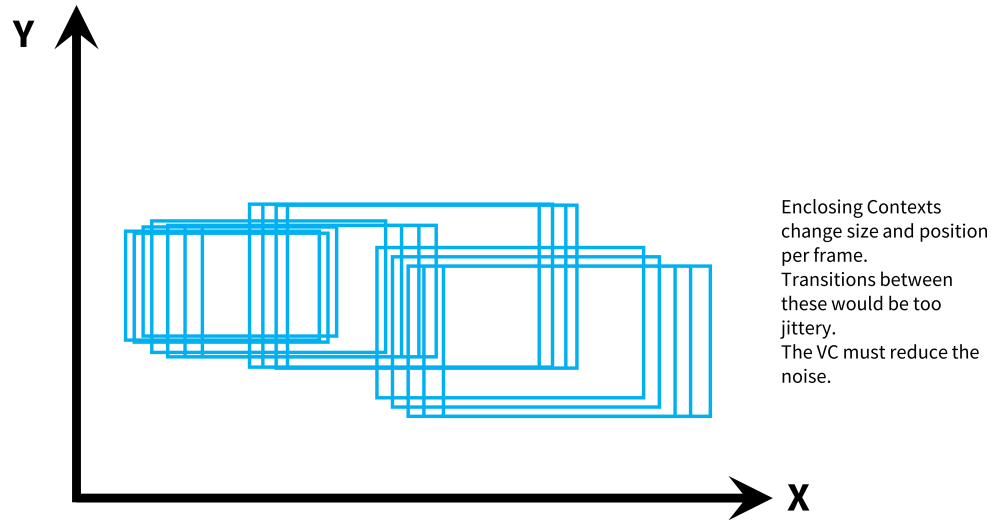


FIGURE 3.24: The presenter's movements change the size and position of the presenter's bounding box

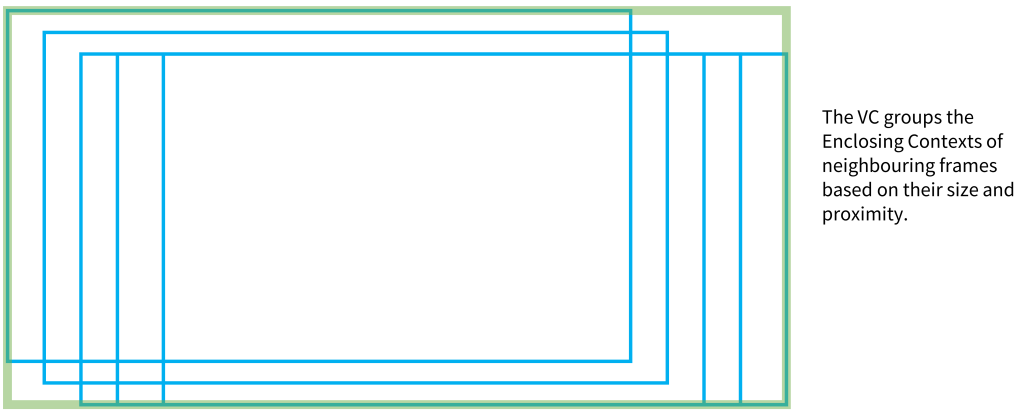


FIGURE 3.25: Contexts are created from separating the bounding boxes into chunks over the length of the video

The pseudocode in Figure 3.26 shows how we used a one-dimensional matrix to store the height values of each frame's enclosing context by writing it as a greyscale pixel where the heights were normalised to range from 0 and 255 (where 0 is the smallest recorded height and 255 is the largest). We then blurred the pixels to smooth out any noise before applying the Canny edge detection algorithm [37], [54] to find edges in the matrix which we could use to create distinct height groups which we called 'chunks' in part 2 as shown in Figure 3.27.

In Figure 3.28 we see that the cropping window changes its height constantly as a result of the minor adjustments to position and scale over time. There are visibly significant differences in average height over time when looking at the graph, but this is not obvious when looking at the heights in an array. The VC uses the Canny Edge detector to find these points of significant change as shown in Figure 3.29.

```
1.  maxHeight = 0
2.  FOR i IN range( contextFrames.size() ):
3.      IF contextFrames[i].enclosingContext.height > maxHeight:
4.          maxHeight = contextFrames[i].enclosingContext.height
5.  matrixOfHeights(1, contextFrames.size(), CV_64FC1) #use only 1 greyscale channel for 1 X n matrix
6.  FOR j IN contextFrames.size():
7.      matrixofHeights[j] = (contextFrames[j].enclosingContext.height / maxHeight) * 255
8.  kernelSize = 3
9.  blur(matrixOfHeights, Size(1, 9))
10. Canny(matrixOfHeights, kernelSize)
11. Vector<int> edges = matrixOfHeights
```

FIGURE 3.26: generateZoomLevels Part 1 - Use a linear matrix and the Canny edge detector to find zoom level chunks

```
12. chunkID = 0
13. chunkStart = 0
14. multiPresenterCount = 0
15. maximumContextHeight = 0
16. sumContextHeight = 0
17. heightCounter = 0
18. chunks = contextFramesManager.chunks
19. FOR z in ContextFrames.size():
20.     IF contextFrames[z].enclosingContext.height > maximumContextHeight:
21.         maximumContextHeight = contextFrames[z].enclosingContext.height
22.  sumContextHeight += contextFrames[z].enclosingContext.height
23.  heightCounter ++
23.     IF edges[z] > 0:
24.         contextFramesManager.averageContextHeights.push_back(sumContextHeight / heightCounter)
25.         sumContextHeight, heightCounter = 0
26.         chunks.emplace_back(Chunk(chunkID, chunkStart, z, Chunk::Zoom, multiPresenterCount))
27.         chunkStart = z + 1
28.         chunkID ++
29.         multiPresenterCount = 0
30.     IF contextFrames[z].multiPresenter:
31.         multiPresenterCount ++
```

FIGURE 3.27: generateZoomLevels Part 2 - Set the enclosing context height to the maximum height measured in the chunk

After normalising the heights between 0 and 255 the VC smooths the heights to bring similar height values closer together such that significant changes are made more obvious. A one-dimensional edge detector is then run over these smoothed heights to determine where the boundaries for the video chunks should be. A Chunk object is then stored for each interval found by the edge detector where the start frame, end frame, motion type, and other important items of information are recorded. The VC also sets the height of each frame in a chunk to be the average height of the enclosing contexts in the chunk. This process is repeated for each chunk in succession and all chunks are saved for use in the next stage of the process. After the heights are adjusted, the VC modifies the width of each frame such that a consistent aspect ratio is maintained throughout the video.

Lecturer tracking and framing rule 2.5 in [23] recommends only moving the camera when the presenter leaves a specified zone or zooming out when smooth tracking is not possible. Our method of adjusting the zoom levels not only follows this heuristic but also follows rule 2.4 which states that context is required to determine which action the camera operator needs to perform. By dividing the video into distinct chunks, the VC makes contexts for which it can make separate framing choices and each context can be joined with a transition which comprises panning and zooming at the same time.

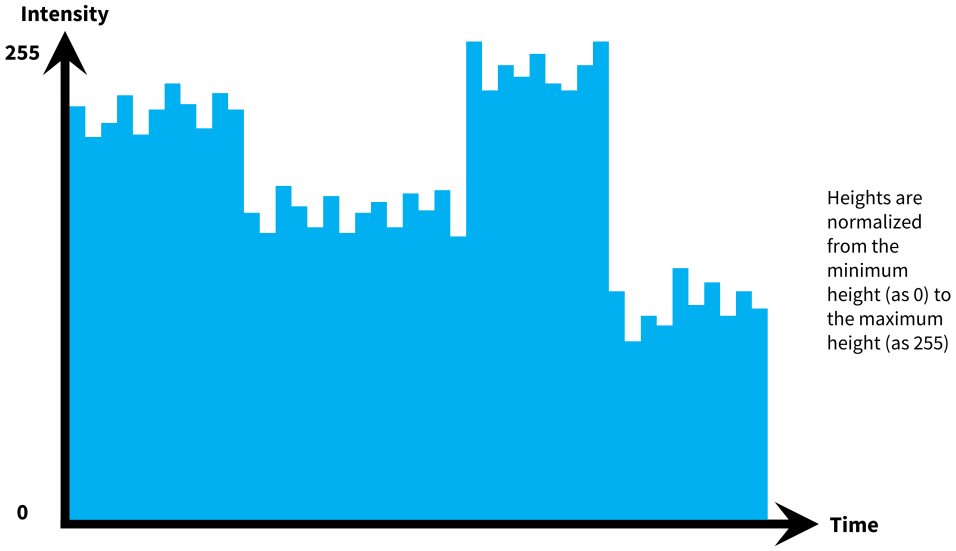


FIGURE 3.28: Example of cropping window heights in a video before smoothing and edge detection

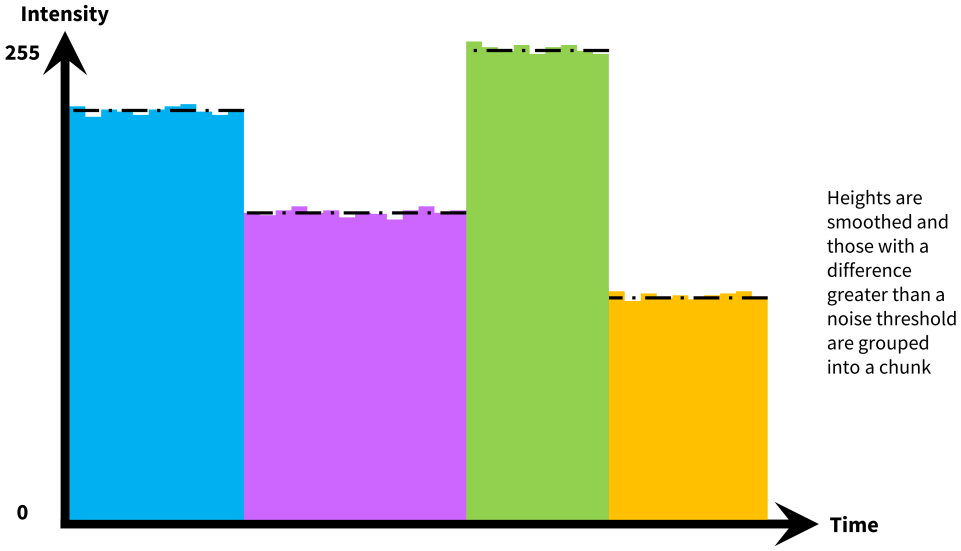


FIGURE 3.29: Example of cropping window heights in a video before smoothing and edge detection

3.2.7 Create the camera transitions

The last step in the Analysis.run() method is to generate the transitions which will determine the final position of the cropping window on each frame of the video so the information can be recorded in the output file.

The VC begins this process by iterating through all the frames and comparing how the Enclosing Contexts in neighbouring frames differ in position.
This gives the VC an idea of the motion in the venue.
Once the VC has all the motion between each Enclosing Context, it must combine all the motion in the same direction so that it becomes only one transition.

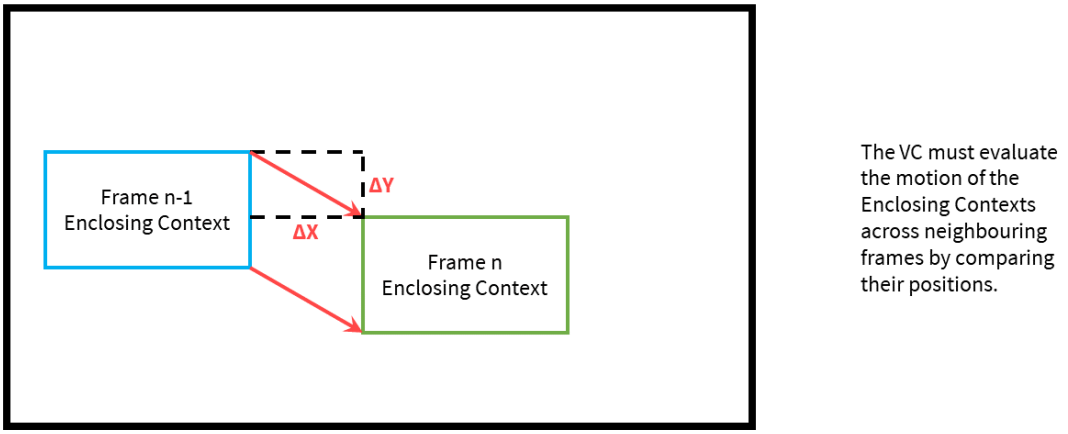


FIGURE 3.30: The difference in motion is calculated between frames to determine the start and end of transitions

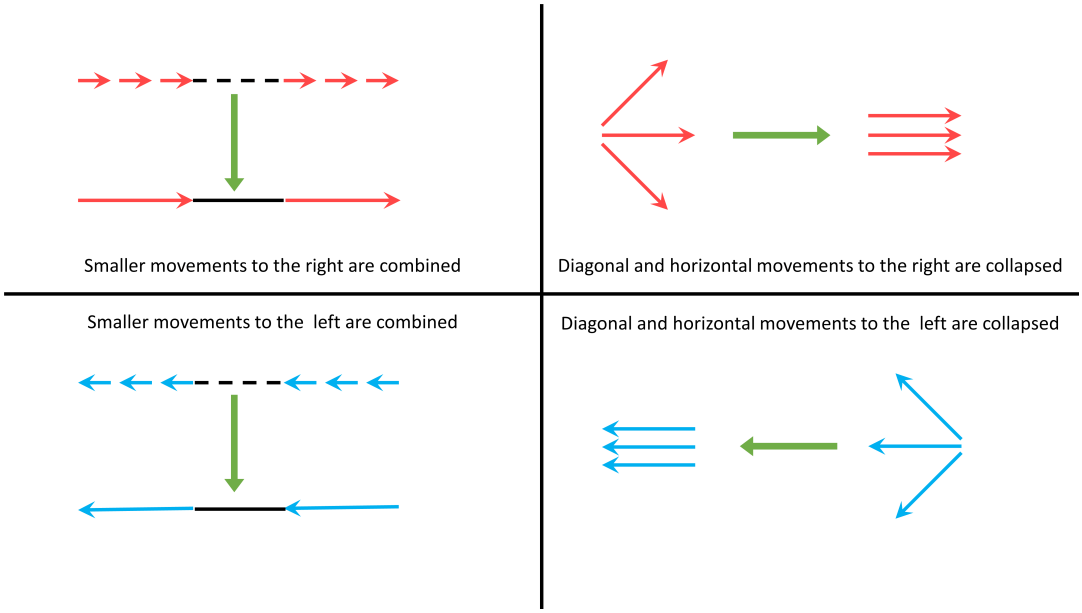


FIGURE 3.31: Successive movements in the same direction are combined into a single transition

Figure 3.31 illustrates how the VC combines similar movements between frames.
All successive movements going the same direction are combined until a move in the opposite direction (or a static video segment) is reached.

If the difference between 2 neighbouring frames' enclosing contexts is smaller than a threshold value, it is considered a static difference and no motion is recorded. This threshold is calculated using half the width of the current frame's Enclosing Context and a constant which we adjusted as part of the user evaluation, called the *transitionLazinessFactor*. This constant, ranging from 0 to 1, influences how responsive the VC is to motion in videos and works as a noise filter for the motion between frames.
For a value of *0*, only movements shorter than the average width of the presenter's bounding box are converted to static video segments. This makes the VC sensitive to small movements and causes the output video to have many smaller transitions.
A value of *1* maximises the noise filtering effect by only recording a difference greater than half the width of the current Enclosing Context as a move. This causes the output video to have fewer transitions and adjustments occur only when major changes in position are detected.

Since the VC already determined the zoom levels in the previous stage, all height changes are ignored, so only the horizontal movements are taken from diagonal movements.

After the motion is identified, the VC creates transitions by animating the enclosing contexts using accelerating and decelerating movements to mimic the behaviour of a camera moved by a human camera operator. It does this by calculating the distance from the start point to the endpoint and then distributing the distance covered per frame according to the curve shown in Equation 3.1.

$$\cos(x + \pi) + 1 \; \textit{in domain} \; [0, 2\pi] \tag{3.1}$$

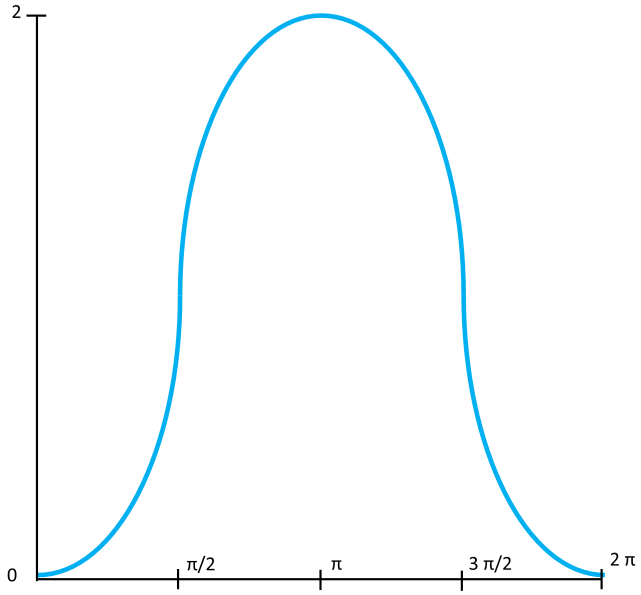


FIGURE 3.32: Cosine Function translated to form a bell-shaped curve for smooth acceleration and deceleration

To reduce the number of transitions, the VC combines panning and zooming into a single operation by creating separate transitions for the vertical and horizontal movements of both the top-left and bottom-right vertices of the enclosing context (see Figure 3.33).

Since each of the four movements has its own acceleration and deceleration, the VC is able to perform the pan and zoom operations simultaneously without causing image distortions or affecting the aspect ratio (as long as the enclosing contexts at the start and end have the same aspect ratio).

The pseudocode in Figures 3.34 and 3.35 show how the VC animates the transition by adjusting the distance covered between each frame in the transition according to the function shown in Equation 3.1.

The VC first checks if the chunk passed in is a static chunk since this method is called for each chunk separately to make the program more modular. In the case of a static chunk, the VC merely returns the enclosing context at the start and the output for the whole chunk will use that rectangle for the cropping window.

If there is motion then the VC calculates the distances between the enclosing context at the start and the enclosing context at the end for the x and y coordinates of the top-left and bottom-right vertices. These distances are then spaced equally for each frame in the chunk in order to normalise the transition.

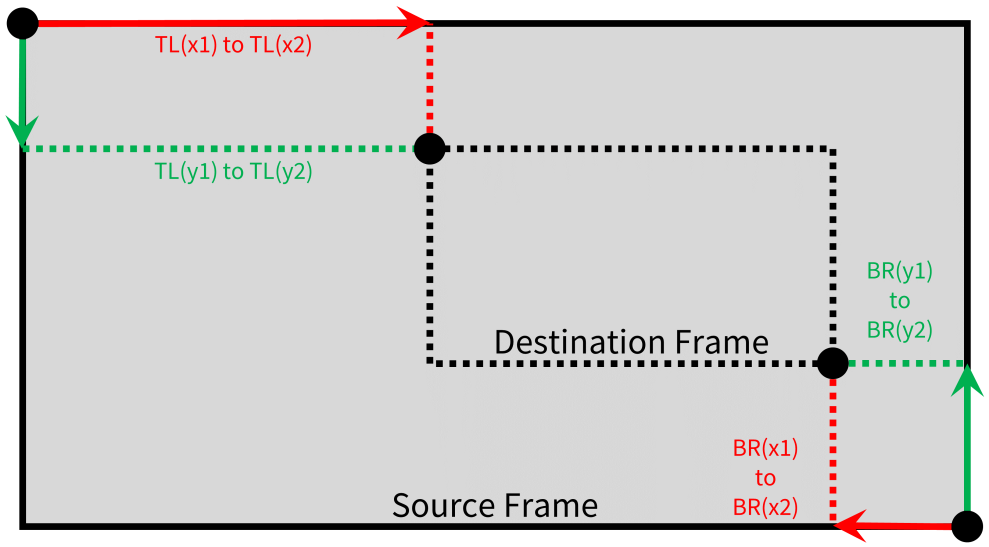


FIGURE 3.33: The transition is applied vertically and horizontally for the top-left and bottom-right vertices

```
1.  frameDifference          =          (endIndex - startIndex) + 1
2.  IF NOT isMoving:
3.      globalCount1 ++
4.      FOR i IN range(frameDifference):
5.          outputFrames.push_back(outputPair(startRect, false))
6.      RETURN startRect
7.  ELSE:
8.      globalCount2 ++
9.      TL_x_diff            =          end.tl().x – start.tl().x
10.     TL_y_diff            =          end.tl().y – start.tl().y
11.     BR_x_diff            =          end.br().x – start.br().x
12.     BR_y_diff            =          end.br().y – start.br().y
13.     TL_x_norm            =          TL_x_diff / frameDifference
14.     TL_y_norm            =          TL_y_diff / frameDifference
15.     BR_x_norm            =          BR_x_diff / frameDifference
16.     BR_y_norm            =          BR_y_diff / frameDifference
17.     domain               =          (2 * CV_PI) / frameDifference
18.     currentRect          =          startRect
```

FIGURE 3.34: Part 1 of the create transitions method

As mentioned in Equation 3.1, the domain is restricted to the interval $[0, 2\pi]$ between transitions in order to stop the cropping window by setting the speed to zero at the start and end of each chunk. This domain is then split into discrete steps by dividing it by the number of frames in the chunk. By doing this, the VC can calculate the exact value in the cosine function according a frame's position in the domain.

```
19. FOR j IN range(frameDiffernce):
20.     TL_x_step       =       TL_x_norm  * (cos((j * domain) – CV_PI) + 1)
21.     TL_y_step       =       TL_y_norm  * (cos((j * domain) – CV_PI) + 1)
22.     BR_x_step       =       BR_x_norm * (cos((j * domain) – CV_PI) + 1)
23.     BR_y_step       =       BR_y_norm * (cos((j * domain) – CV_PI) + 1)

24.     newHeight       =       (contextFrames[j].enclsoingContext.br().y + BR_y_step)
                              – (contextFrames[j].enclsoingContext.tl().y  + TL_y_step)

25.     newWidth        =       (contextFrames[j].enclsoingContext.br().x + BR_x_step)
                              – (contextFrames[j].enclsoingContext.tl().x  + TL_x_step)

26.     currentRect = newRect(newPoint(currentRect.tl().x + TL_x_step, currentRect.tl().y + TL_y_step)
                              , newPoint(currentRect.br().x + BR_x_step, currentRect.br().y + BR_y_step))
27.     outputFrames.push_back(outputPair(currentRect, TRUE))
28. RETURN currentRect
```

FIGURE 3.35: Part 2 of the create transitions method

The VC iterates over each frame in the chunk and multiplies the normalised distance by the cosine function according to its position in the domain. This gives the VC 4 values which it can use to step the enclosing context towards the end positions.

New width and height values are calculated for each frame's cropping window in the chunk by calculating the difference between the top-left and bottom-right vertices on the x and y axes. The cropping windows are then saved as the output frames for the chunk and the VC ends the method by returning the current rectangle (for debugging purposes).

3.3 VC system output

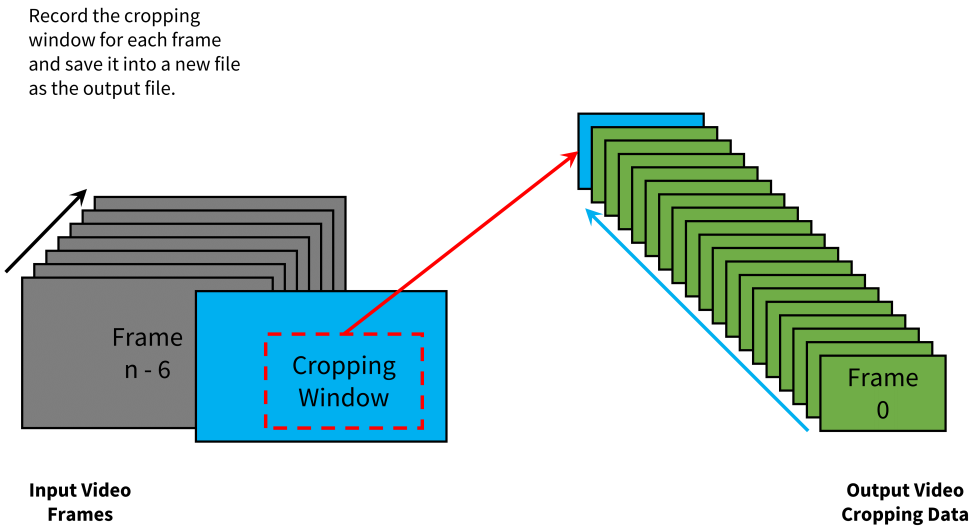


FIGURE 3.36: Each frame in the original video is cropped using the cropping window and the result is saved into the same position in the output video

Once the VC has completed processing, it produces a JSON file containing the cropping rectangle for each frame of the video. The resultant video can be saved as a new video file which can be distributed over the network for viewing. We chose not to crop the footage using the VC directly to dedicate more development effort towards making the VC more adaptable to different network setups.

Some web-based learning institutions use a server to process lecture videos and make them available to students. By making the VC write the information to a data file like the JSON format, these servers could crop the videos on a separate virtual machine to the one running the VC which would speed up the process and make videos available to students faster. It also reduces the need for dependencies between virtual machines on the server. The VC and its output is decoupled such that the only virtual machines on the server that need to have the VC installed are the ones that are running the VC to process new recordings. Virtual machines can be dedicated to cropping the queued videos using only a video processing tool without needing to have the VC installed or even knowing about the VC.

3.3.1 Chapter Summary

This chapter covered the framework of the Virtual Cinematographer in detail in terms of the following: The VC's Input comes from an external module in the form of a JSON file and it contains all the labelled information the VC needs to perform it's operations.
The VC's Internal Processes take this input information and build an internal representation, perform scene analysis, make framing decisions, and calculate transitions between shots.
The VC's Output is written as a JSON file containing the cropping information for each frame. This information is can then be used by an external framework to crop each frame of the input video and write it out as a new video file.
The VC is a module in a larger system which makes it more readily integrated with external programs. It is also able to be distributed on a server such that virtual machines on the server can be dedicated to running only the VC while different virtual machines run the other modules of the system.

Chapter 4

Methodology for User Evaluation

This chapter discusses some of the choices made during the setup phase of the evaluation. We also discuss how the user evaluation was conducted in terms of the experiment setup, and how the data was acquired from the experiment.

4.1 Experiment design

4.1.1 User evaluation layout

In order to test the success of the VC, we needed to do a user evaluation where participants provided a subjective response to the survey of the videos produced by the VC. To achieve this, we needed to create a set of video clips which participants could judge to provide us with the data required to measure the performance of the VC. We designed the experiment to display two video clips, positioned side by side, to participants. By placing the videos next to each other in pairs, participants are able to notice the differences in the videos more easily. The videos in a pair are from the same lesson and time interval in the source footage, but with a different configuration of the VC's settings.

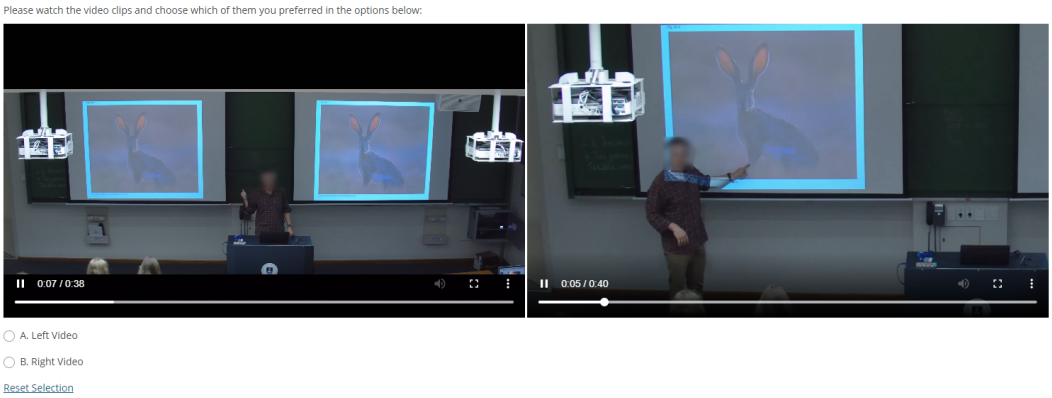
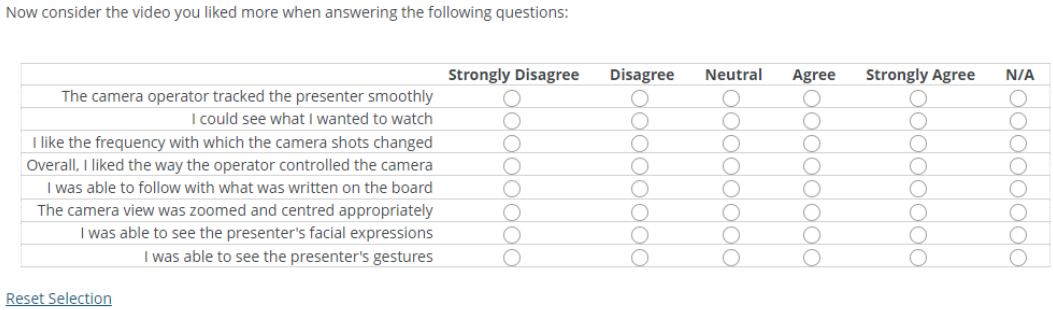


FIGURE 4.1: Screenshot from the user evaluation showing the video pairs arranged side by side with the selection options listed below.

These video players allow participants to play, pause, view in fullscreen mode, and view each of the video clips independently of the other. Participants were asked to play each of the videos and choose which one they preferred. Participants made their choice by selecting the 'Left Video' or 'Right Video' options listed beneath the video players. The videos were set up to play on a loop so participants could review the videos until they were ready to make their choice. A series of questions followed up on this choice to identify possible reasons for their video choice.


Now consider the video you liked more when answering the following questions:

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	N/A
The camera operator tracked the presenter smoothly	○	○	○	○	○	○
I could see what I wanted to watch	○	○	○	○	○	○
I like the frequency with which the camera shots changed	○	○	○	○	○	○
Overall, I liked the way the operator controlled the camera	○	○	○	○	○	○
I was able to follow with what was written on the board	○	○	○	○	○	○
The camera view was zoomed and centred appropriately	○	○	○	○	○	○
I was able to see the presenter's facial expressions	○	○	○	○	○	○
I was able to see the presenter's gestures	○	○	○	○	○	○

Reset Selection

FIGURE 4.2: Screenshot from the user evaluation showing the matrix of questions following each video choice

We asked participants the same eight follow-up questions for every choice they made. Our questions were presented on a five-point scale ranging from 'Strongly Disagree' to 'Strongly Agree' and were stated as affirmations (to which the viewer will respond using the scale). The five-point scale format adds a layer of quantity to this otherwise quality-based evaluation. It also provides us with categorical data which makes it easier to do statistical analyses such as Chi-Square and Cramer's V tests to determine the significance of our findings and the association between various aspects of the VC (if any). We discuss this in more detail in the next chapter.

The questions were taken from questionnaires in published papers on this topic [16], [17]. Rui, Gupta, and Grudin [21] also had a questionnaire in their study. Most of these questions covered the same aspects as the studies done in the aforementioned papers, but the questions were summarised into short phrases rather than full affirmations. In the studies by Rui, He, Gupta, et al. [17] and Rui, Gupta, and Grudin [21] there are questions about how the camera changed to include the audience based on contextual cues. We did not include these questions in our evaluation of the VC since the input videos given to the tracking module (and the resultant tracking information given to the VC) are recorded in venues that are not equipped with audience-facing cameras and we have not included such functionality in the VC. Table 1 in a study by Odhabi and Nicks-Mccaleb [40] shows the usefulness of various teaching aids (e.g. writing on boards, the content shown on projectors, presenter's body language, etc.) being recorded for future revision by students (both before and after editing). While this table might be insightful as to what is most appreciated from these teaching aids, it does not add to the questions we could ask participants in our evaluation of the VC.

We chose a fixed set of questions from the aforementioned papers rather than allowing participants to give long-form answers to questions based on the VC's behaviour to avoid any priming that might result from how the questions were phrased as well as to link the responses to something about the VC directly. Long-form questions are also open-ended and broaden the possible answers given by students rather than narrowing them.

Figure 4.2 shows how the participant was presented with the questions in the evaluation. Table 4.1 shows how these questions linked to the aspects of the VC we wanted to test. Some questions were chosen to get an idea of the participants' opinions on the more general aspects of the video as a whole while others were aimed directly at one specific aspect of the VC when creating the videos. We see that Question 4 was the most general of all where it made an affirmation about the 'overall' camera control. The term 'overall' here is used as a catch-all term and allows participants more freedom to interpret the success of the VC while being specific enough to link it to the way the VC chose to frame the video.

Participants in the user evaluation were volunteers who responded to our invitation to take part in the study. The experiment was hosted on an online platform such that participants were not required to be physically present in a fixed venue. They were informed that their data would remain anonymous and that they were allowed to stop at any point during the evaluation if they felt uncomfortable with continuing. They were also informed that they did not have to provide a reason for their leaving the experiment.

Question Number	Question Affirmation	VC Behaviour to which the question is linkeed
1	The camera operator tracked the presenter smoothly	VC's smoothness of motion during transitions
2	I could see what I wanted to watch	General VC heuristic of 'guiding the viewer's attention to what is important'
3	I like the frequency with which the camera shots changed	Transition frequency during the video
4	Overall, I liked the way the operator controlled the camera	General metric for the VC's control of the framing in the video
5	I was able to follow with what was written on the board	Similar to Question 2, but specific to boards and board content
6	The camera view was zoomed and centred appropriately	More specific than Question 2 and relating to framing and camera zoom
7	I was able to see the presenter's facial expressions	Facial expressions help communicate information to the viewer
8	I was able to see the presenter's gestures	Gives viewers context between what is being said and what is on the boards

TABLE 4.1: Link between questions and the VC's behaviours

There was no time limit assigned to the evaluation, so participants could take their time during the evaluation, but the sum of all the viewing times of the video clips was limited to less than an hour to make it easier for participants and to prevent them from becoming fatigued and answering the questions with reduced concentration. The hosting platform of the evaluation also allowed participants to save their progress and return at a later stage (if they so desired).

While developing the user evaluation, we conducted several pilot tests to refine the interface and make the process easy. Some of the participants complained that the evaluation was too long and it was tiring to go through so many video comparisons. As a result we added in the time limit to the sum of all the video clip times to reduce the strain on participants and minimise the frequency of these complaints. This complaint appeared even during the final evaluation where a participant elected to discontinue the evaluation because it was taking up too much time.

We created variations in the videos by changing the configuration of the VC such that we had video clips from each configuration within the same timestamps of the original video. We used three configurations for the VC in the evaluation: 'Configuration 1', 'Configuration 2', and 'Control'. We limited the number of configurations to three to reduce the complexity of the analysis while testing the limits of the VC when the values are close to the extremes.

The variable which was changed in each VC configuration is called *transitionLazinessFactor*. This variable ranges from 0.0 to 1.0 and determines the pacing range for presenters in the video. If the *transitionLazinessFactor* is set to a value of 1.0, then presenters will pace right up to the edge of the cropping window before the VC makes a transition to adjust the framing. If it is set to a value of 0.0, then the VC will be much more sensitive to changes in the presenter's movements by adjusting the cropping window for any horizontal movements wider than the average width of the presenter's bounding box (to counteract any noise in a stationary presenter's bounding box). Values close to 0.0 are called 'low' and values close to 1.0 are called 'high'. For the user evaluation we use a *transitionLazinessFactor* of 0.1 for Configuration 1, a *transitionLazinessFactor* of 0.8 for Configuration 2, and for Control videos, we left the video unprocessed by the VC and only reduced the output resolution to 1280×720p to match the rest of the video clips.

The values 0.1 and 0.8 were chosen because they were close to the extremes of the *transitionLazinessFactor* variable's range without being exactly on the limits. This keeps the VC framing decisions made by the VC from becoming too uncomfortable for participants while keeping it out of the more stable settings close to the middle of the range.

We discuss the VC's configurations in more detail below.

The Control configuration did not use the VC at all. Video clips using the control configuration were simply taken directly from the source video (including privacy masks), rewritten at a lower resolution, and shown next to the other VC configurations in the evaluation. This was done to determine if presenters preferred an unedited video with a wide-angle view of the lecture venue to videos with VC intervention.

Configuration 1 used a low *transitionLazinessFactor* of 0.1 to make the VC highly sensitive to the presenter's movements while not being so jittery that it would cause discomfort to participants. With this setting, the VC would make adjustments more frequently and for reasons that may not appear obvious to the viewer. We chose this setting to test how a sensitive configuration would be received by participants.

Configuration 2 used a high *transitionLazinessFactor* of 0.8 to make the VC only adjust the cropping window when the presenter came near to the border of the current framing. We chose 0.8 rather than 1.0 to avoid making it look like the VC is unaware of the presenter's coming close to the border and to make it seem (slightly) more responsive.

We also used videos from multiple venues to show participants how the VC responds in different circumstances and with different settings since a single venue layout may not have prompted all the ways the VC can respond to the changes in an environment. We discuss the venue layouts in a dedicated section below. We needed video clips from each venue and, from those venue groups, we needed all the configurations of the VC to be covered. We took the same video segments from each of the source videos (with different venue layouts) and processed them using the different VC configurations (and used them unaltered for the Control).

The video pairs were constructed such that, for each venue layout, all the configurations were placed next to one another.

		Right Video		
		Configuration 1	Configuration 2	Control
Left Video	Configuration 1	N/A	1a	2a
	Configuration 2	1b	N/A	3a
	Control	2b	3b	N/A

TABLE 4.2: Possible constructions for video pairs

In Table 4.2 we see the possible arrangements for video pairs in the user evaluation. The grey blocks are arrangements where the same VC configuration is being compared in both videos of the pair and we have not used those arrangements in the evaluation.

In Table 4.3 we show how we distributed the video pair types. The numbering in each cell indicates which cells are mirror images. It is possible for a video pair to have (for example) a clip from Configuration 2 on the left and a clip from Configuration 1 on the right and this would count as a separate layout to the pair where the clip from Configuration 1 is on the left and the clip from Configuration 2 is on the right. In both layouts, the same Configurations are being compared (not the same clips) but the ordering is different, which is why they are given the same number and the letter changes (symmetric pairing). The ordering of the videos in the pairs was shuffled over the evaluation such that participants could not get used to a pattern. We shuffled the layout of the evaluation template and did not reshuffle the layout for each participant.

We only used 1 video pair including the control configuration for each venue type to focus our strict budget for video clip time to test the VC's configurations. The control configuration was included as a means to test if participants wanted the VC's intervention in the first place and 4 video pairs were enough for this purpose.

	1a	1b	2a	2b	3a	3b
Venue 1	3	1	0	1	0	0
Venue 2	3	1	0	0	1	0
Venue 3	2	2	0	0	1	0
Venue 4	2	2	1	0	0	0

TABLE 4.3: Distribution of video pair types in evaluation

We realised too late that we had included 2 instances of video pair '3a' rather than having one of them as video pair '3b' which is why the column for '3b' is empty. Changing the layout of the evaluation would require a new set of results and we did not have the time or participants to make this correction.

4.1.2 Venue types and their differing layouts:

We chose four venues to use in the evaluation, labelled as 'Venue 1', 'Venue 2', 'Venue 3', and 'Venue 4'. Each venue had different challenges for the tracking module and the VC. We could not choose more than these four layouts because it would require many more video clips and video pairs to be shown in the user evaluation thus making the evaluation too long for volunteering participants. The four venue layouts chosen in the evaluation are all close to the typical lecture venue layout but they each have a separate challenge for the VC's decision-making. This makes them a good test for the VC's output in the evaluation.

We discuss each of the venues in more detail below.

FIGURE 4.3: Screenshot from the video recorded in Venue 1

We see from Figure 4.3 that the lecture venue has two projector screens, a wide space in which presenters may pace, two large projectors (which affect the camera's contrast and lighting), and a few students in the front row (which might cause the tracking module some difficulty at times). All of these challenges, while they affect the tracking module more directly, have an impact on how the VC makes its decisions on how best to frame the video.

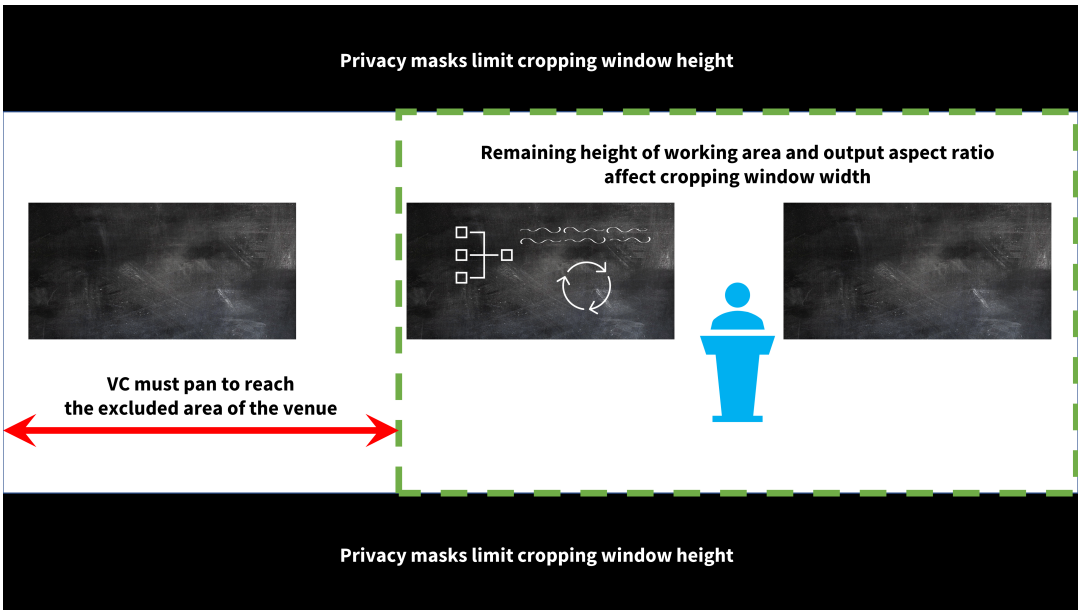


FIGURE 4.4: The effect of privacy masks on the maximum zoom level

Since the height of the working area is reduced by the privacy masks, the VC's cropping window can only be as tall as the height of the working area, which reduces the maximum width to the aspect ratio complement of the working area's height. To include the regions not covered by the maximum zoom level, the VC must pan across the venue.

This reduction in the size of the maximum zoom level means the VC will need to perform more transitions overall. It also introduces a possible scenario where the most important board is on one side of the venue and the presenter is on the other side.

In such a case the VC will choose the presenter over the board since the facial expressions and gestures made by the presenter are more likely to change than the features on the board are over this time.

Venue 2 is similar to Venue 1, except the bottom privacy mask is larger due to a larger portion of the camera view including students who attended the lecture. The content on the projector screens in this lesson is also particularly small, which makes an incorrect framing choice more frustrating for participants in the evaluation. The students who are still visible in this video moved around a lot more than those in Venue 1, which caused more difficulty for the tracking module and indirectly influenced the VC's decisions.

There is also a large area in the front of this venue for presenters to pace. Since the projector screens are down in this video, the blackboards are not in use so the Tracking Module's output regarding board usage is different. This difference, while minor, affects the framing choices made by the VC.

The projector screens are bigger and brighter than the boards in the venue. They also have much finer features on them compared to writing on the boards. This makes it harder for the tracking module to detect writing on these screens and therefore makes it harder for the VC to make framing decisions about it.
The projector screens in this video also show duplicated information which means the VC does not need to include both screens at the same time but there is no way for the VC to know that the information is duplicated.

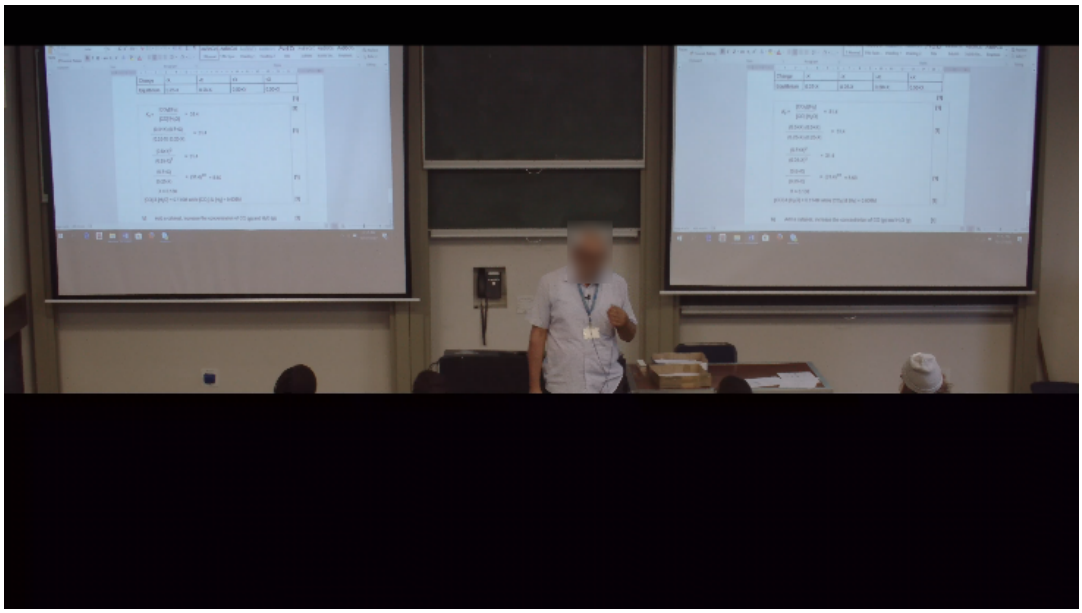


FIGURE 4.5: Screenshot from the video recorded in Venue 2

FIGURE 4.6: Screenshot of Venue 3

Venue 3 has a narrower camera view and, therefore, allows less room for the presenter to pace. There are no projector screens in this video and the privacy masks take up less of the full camera view. Since no students are visible in this video, there were fewer cases where the tracking module could label the venue as "busy" correctly. This means the VC is less likely to zoom out fully and makes a good test case for a video with prolonged close-ups.

The window in the bottom left corner of the venue has bright sunlight streaming through, which affected the contrast in the cameras recording the venue. The presenter in this video also wrote on the boards with a large and clear script, which makes it easier for participants to read the content on the boards and for the Tracking Module to identify the board features correctly. The video recorded in this venue, unlike the others, involves heavy use of the chalkboards. Since the boards slide vertically, this causes both the tracking module and the VC difficulty when trying to maintain continuity regarding board usage. This continuity is not as important because other aspects assist the VC in establishing a context.

Regardless, fluctuating feature counts on the boards complicate the VC's decision-making.

FIGURE 4.7: Screenshot from the video recorded in Venue 4

Venue 4 (Figure 4.7) includes large projectors, but the projector screens are not in use in this video. The privacy masks are thicker at the top of the venue than at the bottom because there is a lot of irrelevant footage in that region of the camera view. Since the VC does not include the regions covered by the privacy masks, this saves the VC from including too much unnecessary footage but it decreases the maximum zoom level of the cropping window for this video.

The camera view in this video is very wide and zoomed out, which compensates for cases when the presenter paces back and forth a lot. In this video, however, the presenter does not move far from the podium in the centre of the venue. By staying close to the podium throughout the video, the presenter's bounding box is combined with the bounding box of the podium in the tracking module's detection algorithm every time the presenter moves behind it. When this happens, the VC assumes something has happened to cause the bounding box to change in size and it zooms out to enclose a larger context. The presenter does not interact with the chalkboards in this video, making the VC's framing decisions more focused on the presenter and the surrounding context.

4.2 Chapter Summary

In this chapter, we detailed the experiment design in terms of the layout for the VC's evaluation and how we obtained data from participants in the evaluation.
We also explained the considerations made when choosing the VC configurations used in the evaluation as well as which venue layouts we used to test the VC's performance in different situations. Our questionnaire was based on questions used in previous research papers in the field and we did not include long-form questions in this evaluation to make interpreting the results easier and more categorical. Participants in the study were volunteers who were all informed before the study that their involvement would remain anonymous and that they were free to leave at any point if they changed their minds.
By choosing to change only the *transitionLazinessFactor* in the evaluation, we simplified the analysis and made it easier to understand how the changes to the VC's configurations affected the video output. We chose four venue layouts in our evaluation for the challenges posed to the VC's framing choices in each environment and how the lecturer (and sometimes students) moved.

Chapter 5

Results, Analysis, and Discussion

This chapter deals with the acquisition of user data and how it was processed before discussing the results obtained from the following aspects:

1. The various configurations of the VC used to decide the framing choices in the output video
2. The different venue layouts used to test the performance of the VC's decision making
3. The experience of the participants in the user evaluation based on their academic background
4. The experience of the participants in terms of the time spent per week watching lecture videos

We discuss these factors in detail in the sections below.

5.1 Evaluation, data acquisition and processing

Once a participant completed the evaluation, their data was recorded on the survey tool. The data included the participant's video choices for each question and their responses to the eight follow-up questions for each video pair.

These submissions were then extracted from the survey tool in Comma Separated Values (CSV) format so the data could be tabulated. The CSV file was reformatted to make it easier to perform statistical analysis using the SPSS program.
We also removed the outliers in the data using SPSS before performing any of the data analysis algorithms detailed below. All entries with missing or unusable data were removed from the database. By the end of this process, 55 complete submissions remained.

We performed Chi-square and Cramer's V evaluations on the data using a cross-tabulation for each question where the responses to the question were compared with the VC configurations, the four lecture venues with different layouts, participant background, and participant average weekly lecture viewing time (the aspects mentioned above). The responses to each of the eight questions ranged from 'Strongly Agree' to 'Strongly Disagree' on a five-point scale (as shown in the figure above). Each cross-tabulation compared these five response categories against only one of the factors (e.g. 'Configuration 1', 'Configuration 1', and 'Control' for VC Configuration type) in each of the four aspects of analysis at a time.

The output for each of the cross-tabulations provided dense tables of information which were then reduced to eight tables containing the most important information from all the output tables. Of these tables, four contained the Chi-Square and Cramer's V tests which report on significance and association and the other four contain the percentages denoting the level of agreement for each question across the different VC configurations and lecture venue layouts. To make reading the results easier, the tables containing the percentages of agreement are shown in section A.3 of Appendix A and the stacked bar graphs below represent the summarised findings from these tables.

The colour scheme used in the tables below is used to extract the positive and negative results at a glance. The colours for each column in the tables below are as follows:

- Expected counts < 5 - Red indicates that more than 20% of the cells in the cross-tabulation had expected frequencies/counts less than 5
- Expected counts < 1 - Red indicates that there is at least one cell with an expected count/frequency less than 1
- Chi-Square - Red cells indicate an insignificant result and green cells represent a significant result
- Fisher-Freeman-Halton exact test - Red cells indicate an insignificant result and green cells indicate a significant result
- Significant - Green means yes and red means no
- Cramer's V - Green indicates a significant result and red indicates an insignificant result
- Association - Red cells indicate a low association and green cells indicate a high association.

The colours in the stacked bar graph are denoted in the key alongside each one. It ranges as follows:

- Strongly Agree - Blue
- Agree - Green
- Neutral - Yellow
- Disagree - Orange
- Strongly Disagree - Red

5.2 Overview of the core statistics used in the analysis of the VC's user evaluation

This section explains some of the core statistics used to analyse the VC's user evaluation.

Categorical (or nominal) variables are sets divided into discrete, mutually exclusive, and exhaustive subsets (or categories). These categories often cannot be ordered or ranked in any meaningful way (e.g. Country of origin, flags of the world). Categorical variables are also referred to as qualitative variables due to their place in quantitative analyses and their unrankable subsets. All the data in this evaluation is categorical, which means we can make comparisons using cross-tabulation.

Cross-tabulations are a means of comparing the results from two (or more) categorical variables by means of a contingency table. From this table, and the associated expected frequencies, it is possible to perform a Chi-Square analysis.

Expected frequencies are used when creating a cross-tabulation and when performing a Chi-Square analysis. They represent the frequency one expects to observe an event occurring according to the probability of that event occurring from all the possible events (e.g. rolling three on a die). It is calculated from the probability of the event occurring divided by the full number of observed events (e.g. how many times should we expect to see a three being rolled if the die was rolled 20 times?).

The Chi-Square test is used to determine if there is a correlation between 2 categorical variables by measuring the difference between the measured and expected frequencies to determine how far the measurements have strayed from the expected frequencies.

The Chi-Square test determines the significance of the difference between the measured and expected frequencies, while the Cramer's V test is used to test the association. In order to obtain accurate results from the Chi-Square and Cramer's V tests, it is important that the expected frequencies in the cells of the Cross-tabulation are no less than 5 (cells with expected frequencies less than 1 are especially bad for accuracy). If more than 20% of the cells in a cross-tabulation have expected frequencies less than 5, the Chi-Square results will not be accurate enough to determine if the differences are significant. In this case, we must use the Fisher-Freeman-Halton Exact test to determine the significance of the results.

The Cramer's V test measures the strength of the association between two categorical variables (those being compared in the cross-tabulation and the Chi-Square test). It is measured as the fraction of the Chi-Square result from the total number of samples taken, χ^2/N (e.g. if the Chi-Square result is 1.25 and the number of samples taken is 36 then the Cramer's V is $1.25/36 = 0.03472$). The Cramer's V test result range is $[0, 1]$, where zero represents no association and one represents the perfect association.

Association refers to how dependent two categorical variables are on each other. For example, in a survey asking males and females whether they preferred odd or even numbers, a strong association would mean that the answer to the question "Odd or even?" would depend on which gender is being asked. If the association is weak, then the preference for odd numbers over even (or vice versa) is independent of the participant's gender.

5.3 Results pertaining to the VC's configurations

Table 5.1 summarises the results obtained from the Chi-square test for independence by cross-tabulating the responses to the 8 questions posed to the participants with the configurations of the VC used in the evaluation.

Question Number	Expected counts > 5	Expected counts > 1	Chi-Square	Fisher -Freeman -Halton exact test	Signifi -cant	Cramer's V	Associa -tion
1	3	1	12.116 $p = 0.151$	N/A	No	0.078 $p = 0.151$	N/A
2	6	3	N/A	29.021 $p < 0.001$	Yes	0.121 $p < 0.01$	Low
3	2	0	13.673 $p = 0.91$	N/A	No	0.085 $p = 0.001$	N/A
4	4	1	N/A	37.781 $p < 0.001$	Yes	0.157 $p < 0.001$	Low
5	3	3	N/A	29.306 $p < 0.001$	Yes	0.142 $p < 0.01$	Low
6	3	1	N/A	28.755 $p < 0.001$	Yes	0.182 $p < 0.001$	Low
7	8	1	N/A	18.614 $p < 0.001$	Yes	0.128 $p < 0.05$	Low
8	7	2	N/A	9.504 $p = 0.245$	No	0.096 $p = 0.220$	N/A

TABLE 5.1: Cross-tabulation of question responses and VC configurations

We used SPSS to perform a cross-tabulation between participants' responses to the questions in the evaluation and the different VC configurations used in the evaluation. The full results of this cross-tabulation can be found in Table A.5 of Appendix A. Participants' responses ranged from 'Strongly Agree' to 'Strongly Disagree' in Table A.5 for each of the eight questions and the table is too dense to extract meaningful information easily.

Each row of Table 5.1, therefore, is the summary of the significance indication data for one of the eight questions in the evaluation (e.g. Row 1 is the summary of the significance indication data for Question 1 in the aforementioned cross-tabulation).

In some of the cross-tabulations, we see that more than 20% of the expected frequencies from the cells are less than 5. This impacts the accuracy of the Chi-Square test so, instead of using the Chi-Square result, we use the Fisher-Freeman-Halton exact test to give us a measure of the association. The Cramer's V and Association columns indicate the magnitude of the association for the cross-tabulation (if any).

We see that the cross-tabulations for Questions One, Three, and Eight all have results where the p-value is greater than 0.05. This means we fail to reject the null hypothesis that there is no association in our findings for those cross-tabulations. The remaining rows all have p-values less than 0.05, which means we reject the null hypothesis. We can, therefore, say that those findings are significant and that they suggest the categories of the two variables show an association.

The following figure shows how each of the VC's configurations was received by the evaluation's participants. A more detailed table (and analysis of the results) can be found in Section A.3.1 of Appendix A.

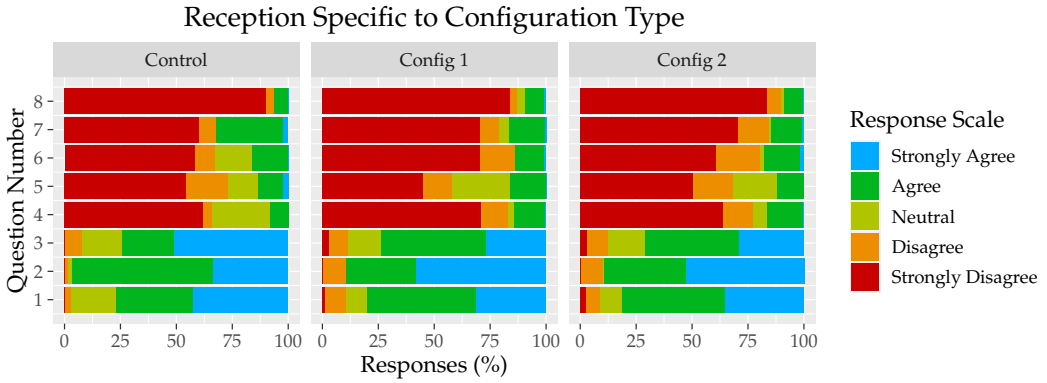


FIGURE 5.1: Responses from participants during the evaluation of the VC pertaining to the VC configuration type

When looking at the results from Figure 5.1, we must note that the findings for Questions 1, 3, and 8 are not significant (as shown in Table 5.1) so we can only make conclusions from the remaining questions' cross-tabulations. Nevertheless, it is worth noting the patterns that can be observed when looking at these findings. We see that Configurations 1 and 2 both have the highest proportion of responses in the 'Agree' category while the Control configuration has the highest proportion of responses in the 'Strongly Agree' category. We observe a shift in the level of agreement between each of the VC's configurations and the Control configuration.

Question 2 asked participants if the VC was showing them what they thought was the most important information for the context. We notice that a higher proportion of participants responded with 'Strongly Agree' for both configurations in comparison to the 'Control' configuration which had a higher proportion of 'Agree' responses. This indicates that both configurations assisted the participants to see what they wanted to watch better than the 'Control' configuration.

Question 4 was intended to give participants a chance to indicate their impressions of the VC without having to use any specific terms or jargon from videography or cinematography. The largest proportion of responses for each configuration is 'Strongly Disagree' with the highest proportion being in 'Configuration 1'. There is a smaller difference between the 'Control' configuration and 'Configuration 2' and we see responses to these configurations are also spread out across 'Agree', 'Disagree', and 'Neutral'.

We must use the other negative responses to indicate why this might be, but future work will also need to expand on the reasons why the participants were not satisfied with the overall camera operation and how it can be improved.

Question 5 asked if participants could follow along with the board content. The largest proportion of responses is in the 'Strongly Disagree' category for all 3 configurations, with Configuration 1 having the smallest proportion of the 3. We interpret this information using the context of the way the VC was configured. The Control configuration shows only the full venue without any post-processing from the VC. Both 'Configuration 1' and 'Configuration 2' have prioritised the presenter over the content on the boards and projectors, so the VC will focus on the presenter when it is forced to choose between the presenter and the content on the boards. The large proportion of 'Strongly Disagree' responses is most likely from this setting in Configurations 1 and 2 and from the zoomed-out view of the Control configuration which makes the writing on the board more difficult to read.

Question 6 asked participants if they thought the camera was 'zoomed and centred appropriately'. We can see from the negative response proportions that the camera view was not zoomed and centred appropriately for any of the configurations. The 'Control' configuration had no VC, which meant the view was centred, but not zoomed in enough for participants to agree with the question. In the case of the other 2 configurations, 'Configuration 1' is adjusting the camera too much and 'Configuration 2' is not adjusting the camera enough to satisfy participants.

If the presenter paces in the video clip, the VC in 'Configuration 2' will zoom out to enclose the area in which the presenter paces. This means the camera view for that section of the lecture would be too zoomed-out for participants to see the presenter's facial expressions. A zoomed-out camera view also includes areas of the lecture venue where nothing important is happening.

'Configuration 1', on the other hand, will zoom in on the presenter. The zoomed-in view means the VC must move around more to keep the pacing presenter in view, which increases the frequency of the VC's camera transitions. This could explain the negative responses since the VC should minimise the number of camera transitions where possible. A larger proportion of responses were 'Strongly Disagree' for this configuration in comparison to 'Configuration 2' and 'Control'. Also, the proportions for 'Strongly Disagree' responses for 'Control' and 'Configuration 2' are very close. Therefore, while 'Configuration 2' and the 'Control' are slightly better at keeping the view zoomed and centred appropriately, there is no configuration from the VC (or the 'Control' configuration) in the evaluation which keeps the camera view centred and zoomed to a satisfactory level.

Question 7 tested if participants could see the presenter's facial expressions. The largest proportion of responses for all three configurations is 'Strongly Disagree' and configurations '1' and '2' have higher proportions than 'Control'. This is likely due to how the VC zooms out in this configuration. The emphasis is to limit transitions as much as possible while including the presenter in the frame at all times. In cases where presenters pace a lot, this results in a zoomed-out framing and this causes the facial expressions to be small and difficult to see. The reception for 'Control' is similar to that of 'Configuration 1' since the footage was not modified at all by the VC in 'Control', so the entire lecture venue is included in the frame. This makes the presenter's facial features the smallest and most difficult to see for the whole lesson due to the maximum zoomed-out view.

'Configuration 1' received a similar reception for a different reason than 'Configuration 2'. The primary focus in 'Configuration 1' is to keep the presenter in view at all times without much interest in minimising the number of camera transitions. This means the VC will stay zoomed in more than it did in other configurations, but it will move around more as well. This movement is likely what is distracting participants to the extent that they are not focusing on the presenter's facial expressions, but further investigation in future work is required to confirm this.

5.4 Results pertaining to the venue types used in the study

The following table summarises the data obtained by cross-tabulating the responses to the 8 questions posed to the participants concerning the types of venues used in the evaluation.

Question Number	Expected counts > 5	Expected counts > 1	Chi-Square	Fisher -Freeman -Halton exact test	Signifi -cant	Cramer's V	Associa -tion
1	2	0	97.646 $p < 0.001$	N/A	Yes	0.180 $p = 0.01$	Low
2	8	4	N/A	79.280 $p < 0.001$	Yes	0.168 $p < 0.001$	Low
3	0	0	68.993 $p < 0.001$	N/A	Yes	0.154 $p < 0.001$	Low
4	4	0	61.508 $p < 0.001$	N/A	Yes	0.163 $p < 0.001$	Low
5	4	4	N/A	182.955 $p < 0.001$	Yes	0.289 $p < 0.001$	Low
6	8	0	N/A	56.565 $p < 0.001$	Yes	0.161 $p < 0.001$	Low
7	8	0	N/A	48.328 $p < 0.001$	Yes	0.174 $p < 0.001$	Low
8	8	2	N/A	54.798 $p < 0.001$	Yes	0.187 $p < 0.001$	Low

TABLE 5.2: Cross-tabulation of question responses and Lecture Venues

As with Table 1, each row represents a cross-tabulation between the responses to the question and the videos recorded in different venues.

From the table above, we see that only rows 1, 3, and 4 have few enough cells with expected frequencies less than 5, thereby making the Chi-Square test reliably accurate. We must, therefore, use the Fisher-Freeman-Halton exact test for the rest of the rows to determine association. We also see that all the cross-tabulations in this table are significant, which means we use the Cramer's V test to evaluate the magnitude of association for each cross-tabulation. The Cramer's V test reveals that all of the questions have a low association. This suggests a low practical significance to how the participants' responses depend on the venue type.

The following figure shows how each of the lesson venues was received by the participants for each question of the evaluation. A more detailed table and analysis can be found in Section A.3.2 of Appendix A.

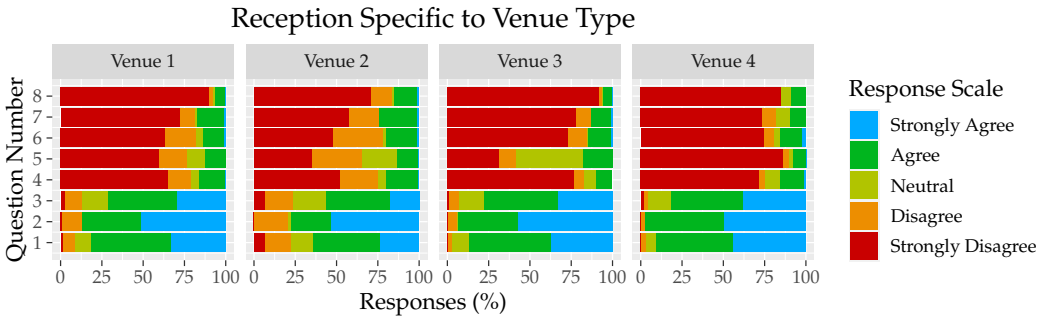


FIGURE 5.2: Responses from participants during the evaluation of the VC pertaining to the VC venue type

From Figure 5.2, we see that Questions 1, 2, and 3 were received well, with each venue having higher proportions of 'Strongly Agree' and 'Agree' responses. The remainder of the questions all have responses skewed towards 'Strongly Disagree' and 'Disagree' for each of the 4 venue types.

For Question 1, the VC performed its operations smoothly in all the venues due to the high levels of agreement per venue for this question. Venue 2 had the lowest proportion of positive responses compared to the other venues. This is because the students in the venue moved around a lot which caused the tracking module to label the scene as busy during those contexts. The VC zooms out whenever the tracking module is labelled busy. If the video has many patches of busy sections such that the VC is forced to change the zoom level too frequently, it could affect the number of positive responses despite the smoothness of those transitions (as we see with Venue 2).

There are high levels of agreement for each venue in Question 2 and, since our findings are significant, we can see that the VC succeeded in its framing choices. The worst of the 4 venues for Question 2 was Venue 2 since the venue had a lot of students moving around near the presenter (as mentioned above). This lesson also made heavy use of the projectors, displaying fine writing which is too difficult to read in the video. This might explain why the proportion of positive responses was so small compared to the other 3 venues.

For Question 3 we must compare the camera transition frequency for each venue. Venue 4 has the most stability in the lesson since the presenter remained close to the centre of the venue and was not interacting with the boards at all. There are also very few disruptions to the tracking of Venue 4 since the students were not in view, which meant the venue was not marked as 'busy' very often.

Venue 3 has the second highest proportions. This venue is smaller than the others and the presenter is moving around and interacting with the boards often which means the VC would be changing shots regularly. These shots were appropriate for the context since participants answered positively about this venue in the previous question, which focused on being able to see the most important information. Therefore, the VC was making camera transitions at the appropriate times and including the most relevant information.

Venue 1 has the second lowest level of agreement among the 4 venues in this question. The presenter in this venue is pacing and gesturing while explaining concepts around the content being displayed on the projector screens. Since there were some students in the front of the venue, the tracking module would mark some sections of the video as 'busy' whenever these students were moving around too much. This caused the VC to transition to another camera view to frame things in a better way. This might be happening too often, which is why we see lower proportions of positive responses.

Venue 2 has the lowest levels of agreement out of the 4 for Question 3. This is, once again, due to the movements of the students at the front of the venue. The VC changed the framing to adjust for the busy scenes too often, which resulted in the poorest participant reception of the 4 venues.

While the association for Question 4 is low, we can assume that the reason for such large levels of disagreement for each of the venues is that the VC did not operate favourably. Since the question is phrased in such a vague way, it leaves room for participants to compare the VC's camera operations to that which they are used to seeing. It could be that students have come to expect lectures that were recorded solely for remote learning. Such videos have a different camera setup and the recording process is different as well, which might explain the poor reception by participants.

Question 5 asked participants about how the VC framed the boards in the lecture recordings. In some cases the VC must choose between framing the presenter using a close-up shot, focusing on the contents of the boards while ignoring the presenter, or finding some intermediate between these extremes. By default, in such a case, the VC chooses to favour presenters and whatever facial expressions and gestures they use to convey their material.

This question is asking participants specifically about the writing on the boards and how well students could follow it. Some of the venues did not use the boards and, instead, included projectors, so students may not have considered projector screens as boards.

It is also important to note that Venue 4 did not use any boards at all. Neither the chalkboards nor the projector screens were used to convey the lecture material in this video, which could also explain why participants would answer negatively about being able to follow the boards. In future work, we could exclude options that do not apply per video pair such that participants' responses are easier to interpret.

Question 6 targets both the zoom level and the focal point of the VC's framing choices. It is phrased such that it does not prime participants or lead them to think in a specific way when answering. The definition of appropriate in the context of the VC is left open to get participants' higher-level impressions of both the zooming and focus of the framing choices made by the VC.

We see that all 4 venues have a similar proportion of negative responses to this question, and all these proportions are high, which indicates that the VC is making framing choices that are not pleasing to participants regardless of the venue or its recording conditions. We see from the cross-tabulation between the responses to Question 6 and the different venues that the participants received all venues poorly, which further indicates that there is something wrong with its framing choices in terms of zoom and focal point for each context and different environments.

Question 7, while being more direct than the previous one, is also phrased to avoid priming participants to think in a specific way. Since Venue 2 did not make use of the boards, and a large portion of the venue was masked off with privacy masks to protect the identity of students in the venue, the VC was able to use a more zoomed-in view of the lesson (even in its loose framing shots). This close-up view of the presenter and the surrounding context might be why the levels of disagreement for this venue are lower compared to the others.

Question 8 asked participants if they could see the presenter's gestures. The venue with the highest levels of disagreement was Venue 2. It could be that the presenter did not gesture as much as the presenters from other venues which made the percentage so high, or the students in the lesson were moving around too much which caused the scene to be labelled 'busy', in which case the VC zooms out to try and accommodate as much of the 'busy' scene as possible. By zooming out, the VC made it difficult for participants to see any gestures the presenter might have been making.

Another reason for the high levels of disagreement could be bad lighting. We see in Venue 1 that the presenter had switched off the lights in the front of the venue to make it easier for students to see the projector screens more clearly. This lack of lighting might be preventing students from being able to see the presenter's gestures.

There is a low association between Question 8 and the venues in which the videos were recorded. This means that participants were disagreeing with Question 8 with no practical significance over which venue the videos were recorded in. More research is required to determine exactly what is causing the high levels of disagreement regarding this question.

5.5 Results pertaining to participant background

The following table summarises the data obtained by cross-tabulating the responses to the 8 questions posed to the participants concerning participant background. Each row in the table represents a cross-tabulation pertaining to a specific question from the user evaluation.

We see that only Question 3 has few enough cells in its cross-tabulation with expected frequencies less than 5, which would affect the Chi-Square test's accuracy. We must therefore use the Fisher-Freeman-Halton exact test for the other questions to determine significance and association. We notice that Questions 4, 6, and 8 have p-values greater than 0.05 which means we fail to reject the null hypothesis that there is no association in our findings for those cross-tabulations.

Question Number	Expected counts > 5	Expected counts > 1	Chi-Square	Fisher -Freeman -Halton exact test	Signifi -cant	Cramer's V	Associa -tion
1	6	1	N/A	57.376 $p < 0.001$	No	0.128 $p < 0.001$	Low
2	11	7	N/A	47.973 $p < 0.001$	Yes	0.112 $p < 0.01$	Low
3	4	1	85.657 $p < 0.001$	N/A	No	0.158 $p < 0.001$	Low
4	8	4	N/A	22.974 $p > 0.05$	Yes	0.096 $p > 0.05$	N/A
5	8	5	N/A	26.185 $p < 0.05$	Yes	0.103 $p > 0.05$	N/A
6	11	2	N/A	21.431 $p > 0.05$	Yes	0.096 $p > 0.05$	N/A
7	12	4	N/A	26.572 $p < 0.05$	Yes	0.109 $p = 0.05$	Low
8	12	5	N/A	15.884 $p > 0.05$	No	0.087 $p > 0.05$	N/A

TABLE 5.3: Cross-tabulation of question responses and participant background

The remaining questions all have p-values below (or equal to) 0.05, which means we reject the null hypothesis. We can, therefore, say that those findings are significant and that they suggest the categories of the two variables show an association.

Figure 5.3 shows how participants responded to each question in the evaluation according to participant background (based on year of study or staff level). A more detailed table and analysis can be found in Section A.3.3 of Appendix A.

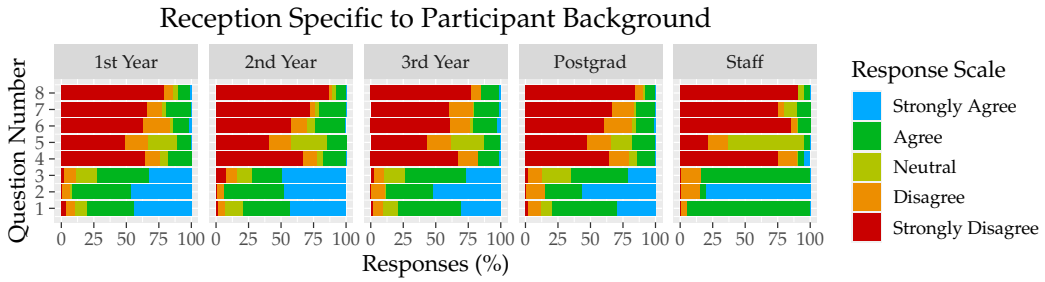


FIGURE 5.3: Results of the VC evaluation pertaining to participant background

We did this comparison in order to judge the weight of their responses in the study according to their experience with watching lecture recordings. It is more likely that a more senior participant (e.g. a third-year student) would have more experience with watching lecture recordings.

Question 1 asked participants if the VC was tracking the presenter smoothly. We notice a larger proportion of participants responded in agreement ('Strongly Agree' and 'Agree') across all participant background groups. We notice that the percentages of 'Strongly Agree' responses decrease as the participant background changes from 1st-year through to staff. We suspect this is due to the increased experience with watching lecture videos; as the year of study (or background) increases, so does the experience with watching lecture videos. This increased experience with lecture videos could be tempering the enthusiasm with which participants responded to the VC's smoothness in tracking the presenter.

Question 2 was phrased, 'I could see what I wanted to watch.' For this question, we notice that the percentage of 'Strongly Agree' responses increases as the categories change from 1st-year to staff which is the inverse of what we observe in Question 1. This could, again, be due to the increased experience in participants with a higher year of study. They know what they want to see in a lecture video and are more satisfied with being able to see it than the less experienced participants.

Question 3 was phrased, 'I like the frequency with which the camera shots changed.' and we notice that there is a high percentage of positive responses for this question across all the background categories. We notice the 'postgrad' category has a large percentage of 'neutral' responses. This question also has the most spread-out set of responses among all 5 categories. We observe that the percentages of positive responses across the 5 categories do not change in a linear fashion as with Questions 1 and 2. Instead, we observe a peak of 'Strongly Agree' responses in the '2nd Year' category and then a decrease as the year of study increases. If we contextualise this observation with experience, we can assume that second-year participants had more confidence in the VC's shot change frequency, while the more experienced categories have less confidence since they have seen more lecture videos before the VC evaluation and, thus, have higher standards.

Question 5 was phrased as follows, 'I was able to follow with what was written on the board.'. For this question, we see that there is a higher proportion of negative responses ('Strongly Disagree' and 'Disagree'). Four of the five categories had large percentages of 'Neutral' responses (the largest proportions of all the questions for this section of the evaluation), which could indicate some participants were unsure about their responses to this question. This might have to do with the videos recorded in 'Venue 4' where no board usage occurred and some participants answered 'Neutral' rather than 'Strongly Disagree'.

Question 7 checked if participants could see the presenter's facial expressions. We note that this question (as with Question 5 above) has a higher percentage of negative responses. We note that all categories have high percentages in the 'Strongly Disagree' category which means that most of the participants were not able to see the presenter's facial expressions. This could be due to the VC's having to zoom out to accommodate more of the lecture environment. Zooming out this much then reduces the size of the presenter's face and, consequently, makes it more difficult to see facial expressions.

5.6 Results pertaining to participant average weekly viewing time

Table 5.4 summarises the data obtained by cross-tabulating the responses to the 8 questions posed to the participants concerning the time participants spend (on average) watching lecture videos per week. Each row in the table represents a cross-tabulation pertaining to a specific question from the user evaluation.

We note that Questions 2, 6, 7, and 8 all have more than 20% of their cells in the cross-tabulation where the expected count is less than 5. This means that, for these cross-tabulations, we must use the Fisher-Freeman-Halton exact test, rather than the Chi-Square test to determine significance and association. We also note that only Questions 1, 3, and 8 have p-values greater than 0.05, which means we fail to reject the null hypothesis for these questions and say that the findings for these questions are not significant.

For the remaining questions, we notice that there is a low association (from 0.097 to 0.179) between the responses to the questions in the evaluation and the number of hours spent per week (on average) watching lecture videos.

Figure 5.4 shows how participants responded to each of the questions in the evaluation grouped according to the average time spent watching lecture videos. A more detailed table and analysis can be found in Section A.3.4 of Appendix A.

In Figure 5.5, we see that most responses came from the 'Postgrad' category but the majority of these participants spent less than two hours per week on average watching lecture videos.

Question Number	Expected counts > 5	Expected counts > 1	Chi-Square	Fisher -Freeman -Halton exact test	Signifi -cant	Cramer's V	Associa -tion
1	3	1	12.116 $p = 0.151$	N/A	No	0.078 $p = 0.151$	N/A
2	6	3	N/A	29.021 $p < 0.001$	Yes	0.121 $p < 0.01$	Low
3	2	0	13.673 $p = 0.91$	N/A	No	0.085 $p = 0.091$	N/A
4	4	1	N/A	37.781 $p < 0.001$	Yes	0.157 $p < 0.001$	Low
5	3	3	N/A	29.306 $p < 0.001$	Yes	0.142 $p < 0.01$	Low
6	3	1	N/A	28.755 $p < 0.001$	Yes	0.182 $p < 0.001$	Low
7	8	1	N/A	18.614 $p < 0.001$	Yes	0.128 $p < 0.05$	Low
8	7	2	N/A	9.504 $p = 0.245$	No	0.096 $p = 0.220$	N/A

TABLE 5.4: Cross-tabulation of question responses and participant average weekly viewing times

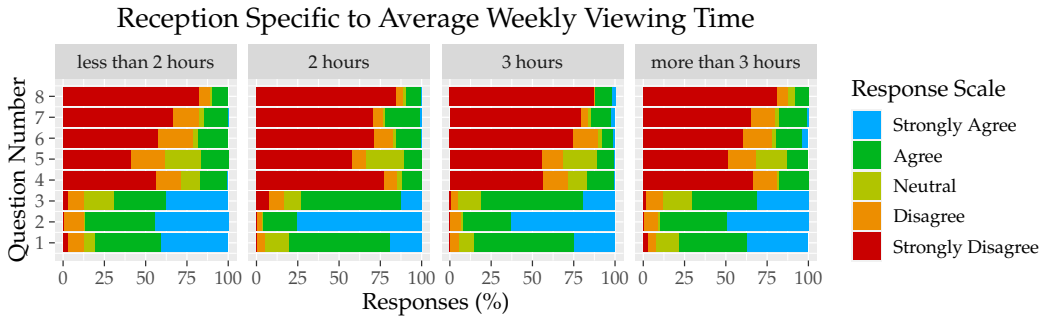


FIGURE 5.4: Results of the VC evaluation pertaining to participant average time spent watching lecture videos per week

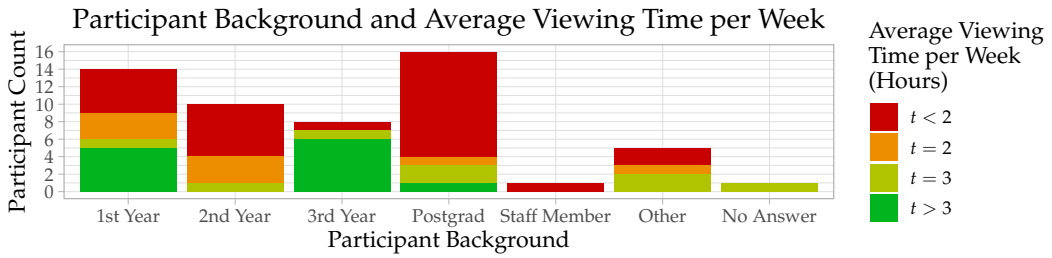


FIGURE 5.5: Proportions of hours spent weekly watching lecture videos according to participant background

We observe that the third-year category has the highest percentage of participants who spend more than three hours per week on average watching lecture recordings. This means that participants from the third-year category have the most experience with lecture recordings. We should, therefore, apply more weight to the third-year responses due to their experience with lecture recordings. Given that there are only 55 participants in total, and the frequencies in the 'background' and 'viewing time' sub-categories are so low, the results may suggest patterns without enough evidence to make conclusive observations.

Question 2 tested if participants were able to see what they wanted to watch when viewing the videos in the evaluation. We observe that there is a higher percentage of positive responses for this question across all the 'viewing time' categories. We observe that the 'Strongly Agree' responses decrease as the weekly viewing time increases. This means that more experienced participants were less confident that they could see what they wanted to watch.

Question 4 tested the overall impressions the VC gave participants about its control of the camera and framing decisions. For this question, we see that there is a higher percentage of negative responses ('Strongly Disagree' and 'Disagree') from all 4 of the 'viewing time' categories. This indicates that even the more experienced participants were not satisfied with the VC's performance in general, which requires more investigation to determine what is causing this dissatisfaction and how it could be improved.

Question 5 tested if participants could follow what was written on the boards. For this question, we notice that there is a higher percentage of negative responses for all 4 'viewing time' categories. We also note that all categories have a notable percentage of 'Neutral' responses. The proportions of 'Strongly Disagree' responses remain the same as experience increases. The VC is not showing viewers what they want to see on the boards since even the more experienced participants responded negatively in a similar proportion to the less experienced participants.

Question 6 was phrased, 'The camera view was zoomed and centred appropriately'. We see that, for this question, there is a higher percentage of negative responses for all 4 categories. The proportions of 'Strongly Disagree' responses increase as participant experience increases, which indicates that the VC's framing choices are not satisfactory in terms of its zoom and centre control. More work is required to determine which of these aspects (zoom or centre framing) is the cause of the negative reception.

Question 7 asked participants if they could see the presenter's facial expressions when watching the videos in the evaluation. For this question, there is a higher percentage of negative responses for all 4 categories. We also see the percentage of 'Agree' responses is higher than that of 'Disagree' except for the 'Less Than 2 Hours' category where the proportions are approximately the same. The proportion of 'Strongly Disagree' responses increases as weekly viewing time increases which indicates that the more experienced participants were less satisfied with how the VC framed the presenter's facial expressions.

5.7 Discussion on analysis of results

Question Number	Question Affirmation
1	The camera operator tracked the presenter smoothly
2	I could see what I wanted to watch
3	I like the frequency with which the camera shots changed
4	Overall, I liked the way the operator controlled the camera
5	I was able to follow with what was written on the board
6	The camera view was zoomed and centred appropriately
7	I was able to see the presenter's facial expressions
8	I was able to see the presenter's gestures

TABLE 5.5: Table showing how each question was phrased as an affirmation

Table 5.5 provides a reference for how the Question affirmations were phrased in the evaluation. Table 5.6 provides a summary of the statistical significance data for each of the four aspects of analysis as well as the VC's reception. Table 5.7 shows the levels of association for each question in the evaluation and for each aspect of analysis (if statistically significant).

	Question Number	1	2	3	4	5	6	7	8
VC Configuration	Significant?	No	Yes	No	Yes	Yes	Yes	Yes	No
	Reception	Pos	Pos	Pos	Neg	Neg	Neg	Neg	Neg
Venue Layout	Significant?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Reception	Pos	Pos	Pos	Neg	Neg	Neg	Neg	Neg
Participant Background	Significant?	No	Yes	No	Yes	Yes	Yes	Yes	No
	Reception	Pos	Pos	Pos	Neg	Neg	Neg	Neg	Neg
Participant Viewing Time	Significant?	No	Yes	No	Yes	Yes	Yes	Yes	No
	Reception	Pos	Pos	Pos	Neg	Neg	Neg	Neg	Neg

TABLE 5.6: Table showing a summary of the findings for the 4 aspects of analysis in terms of statistical significance and participant reception

Question Number	1	2	3	4	5	6	7	8
VC Configuration	N/A	Low	N/A	Low	Low	Low	Low	N/A
Venue Layout	Low	Low	Low	Low	Low	Low	Low	Low
Participant Background	Low	Low	Low	N/A	N/A	N/A	Low	N/A
Participant Viewing Time	N/A	Low	N/A	Low	Low	Low	Low	N/A

TABLE 5.7: Table summarising the levels of association for each aspect of analysis

'Venue Layout' is the analysis aspect where all eight questions have statistically significant results. This means that we cannot make any conclusions from the statistically insignificant results in the other aspects of analysis. Still, we may make observations on the patterns in the data.

When looking at how the VC was received by participants in table 5.6, we notice that the VC obtained a positive reception in Questions 1, 2, and 3 for all aspects of analysis and the remaining questions all obtained a negative reception for all aspects of analysis.
As shown in Table 5.5, the first three questions in the evaluation all relate to the main heuristics of the VC, namely:

1. Smooth camera transitions
2. Guiding the viewer's attention to what is most important
3. Minimising the number of camera transitions

This positive reception is not strongly dependent on any of the aspects of analysis as seen in Table 5.7, however, we cannot conclude this for the grey cells due to a lack of statistical significance. This weak dependence is also observed for the questions with a negative reception.

We will now summarise and interpret the findings for each aspect of analysis below:

The 'VC Configuration' findings show that (despite insignificant data for Questions 1, 3, and 8) the VC was received positively for the main heuristics and negatively for the remaining heuristics. As for the positive reception, the proportions are similar for 'Config 1' and 'Config 2' which are both larger than 'Control'. This proportionality is seen for Questions 1, 2, and 3 and it means that the VC is performing better than the 'Control' configuration for these questions. The VC's configurations are, therefore, better than no intervention for the main heuristics.

The inverse is observed for the remaining questions where 'Config 1' and 'Config 2' have larger proportions of negative responses and, therefore, the VC's configurations are worse than no intervention for these heuristics.

The association score for this aspect of analysis is 'Low' for all questions (except 1, 3, and 8 which are insignificant). This suggests that the VC's configurations are only weakly associated with the VC's reception. We also observe that 'Configuration 1', and 'Configuration 2' have roughly the same proportions in the responses for all the questions. This seems to indicate that, while the VC has outperformed the 'Control' configuration, neither of the configurations has done better than the other.

As 'Venue Layouts', we see that all eight questions are statistically significant and the VC was received positively only for the main heuristics. Since all eight questions have a 'Low' association score, we conclude that the VC succeeded in the main heuristics regardless of the venue layout. Venue 4 had the best positive reception comparatively for the main heuristics and Venue 2 had the smallest proportion of positive responses (as seen in Figure 5.2). In terms of the negative reception, Venue 4 did the worst and Venue 2 had the lowest proportion of negative responses. The questions that received a negative reception were only weakly associated with the layout of the venue. This means that the VC's unsatisfactory performance is more fundamental than the structure of the lecture venue. A weak association is not the same as 'no association', however, and there is a slight influence from the venue's layout on the VC's performance (e.g. Venue 4's lack of board use affecting the responses for Question 5).

Our findings for 'Participant Background' show that all participants, regardless of their background, received the VC well in terms of the main heuristics and responded more negatively to the remaining questions of the evaluation. The 'Low' association score also indicates that the background information of participants has little to do with the way the VC is received by viewers. Nevertheless, we speculate that what little influence the background of the participant has on the reception of the VC is due to the experience (or lack thereof) with watching lecture videos. More experienced viewers have a better understanding of what they want from lecture videos than less experienced viewers, which is why the level of agreement decreases as experience increases (and vice versa for disagreement). The 'Staff' category had the highest proportions of 'Strongly Disagree' responses and we speculate that, as producers of lecture content, these participants are especially critical of how the lecture material is being framed.

The 'Average Weekly Viewing Time' aspect of analysis was intended to give an insight into how increased exposure to lecture videos per week affected the VC's reception in the evaluation. We performed two comparisons in this aspect of analysis:

1. How participants responded to the evaluation questions according to the average number of hours per week they spent watching lecture videos
2. The average viewing time proportions for each background category

This two-pronged investigation helps to link the two aspects of analysis and gives more insight into whether there should be a weight assigned to participant responses based on lecture video exposure and experience. While the 'Postgrad' category has the most participants, the '3rd' Year category has the most exposure and is the second-most in seniority among the 'student' background categories. The 'Postgrad' category had the largest proportion of participants with less than two hours spent on lecture videos per week. When compared with how 'Postgrad' participants responded in the Background aspect of analysis we see that there are more 'Strongly Agree' and 'Strongly Disagree' responses than the less senior students who spent more time per week watching lecture videos. Participants from the 'More than three hours' category have the largest proportions of 'Strongly Disagree' responses and a roughly equal propotion of 'Strongly Disagree' responses to the 'Less than 2 hours' category. This means that, while participants show roughly the same reception to the main heuristics,

the participants with exposure to lecture videos seem more sensitive to the more nuanced heuristics of the VC (like showing facial expressions, board content, and gestures).

A 'Low' association score indicates that the VC's reception is only weakly dependent on the participants' average weekly lecture viewing time. Any influence this aspect has on the VC's reception is most likely linked to a balance between the average weekly exposure to lecture videos and the overall experience with lecture videos in general.

5.8 Chapter summary

In this chapter, we performed a detailed analysis of the results obtained from the VC's user evaluation from four aspects:

1. Participant responses according to the VC's Configurations and the 'Control' configuration.
2. Participant responses according to various venue layouts used in the evaluation.
3. Participant responses analysed in the context of participant background.
4. Participant responses analysed in the context of the average time spent by participants per week watching lecture videos.

Finally, we provide a discussion section for each of these aspects of analysis.

In the next chapter, we conclude our findings and discuss the potential for future work and improvement in this study.

Chapter 6

Conclusions

In this study, we investigated whether a Virtual Cinematographer could be used to reduce the resolution of pre-recorded 4K lecture videos through the use of a frame cropping window. The cropping window is updated according to carefully determined framing heuristics such that the final smaller format video appears to have been recorded by a human camera operator.

Our literature review of the cinematography field suggested that: the VC should track the presenter smoothly so that any transitions have acceleration and deceleration; the footage should also guide the viewer's attention to what is important in the scene while minimising whatever is not relevant to the context; finally, the VC should minimise transitions wherever possible to keep viewers from becoming distracted and disoriented. Other heuristics were used to inform the VC's decision-making, but the ones mentioned above are the most crucial to the VC's successful performance.

We set up our user evaluation to measure the VC's performance for each of the main heuristics (along with some of the minor ones) identified during our review of the literature.

The VC is implemented as the second module in a post-processing system. The first module in the system takes the pre-recorded 4K lecture video and tracks the lecturer (along with other important information) while labelling the data in JSON format as its output. The VC then uses this labelled data to perform scene analysis and decide how best to manipulate the input video cropping such that the output video appears to have been recorded by a human camera operator. The VC always keeps the lecturer in focus for each frame, adjusting the zoom and position of the cropping window according to the surrounding context (e.g. including any boards in use near the presenter). By doing this, the VC implements the heuristic of guiding the viewer's attention to what is important. The VC also divides the video into contextual "chunks" such that it only makes transitions between these chunks and reduces unnecessary adjustments. This ensures the VC only makes transitions when required. When the VC must make a transition, the motion of the cropping window performs the horizontal and vertical pan movements simultaneously with changes in zoom (to minimise the number of transitions) while smoothing the acceleration and deceleration of transitions to satisfy the heuristic of transition smoothness.

A user evaluation was set up to test the VC where we showed participants pairs of adjacent video clips playing simultaneously and on loop for easy comparison. Participants were asked to choose which of the two videos they liked more for each pair and answer eight follow-up questions based on their choice. This way, we could see which videos they preferred in each comparison and which heuristics were being fulfilled in the videos they chose.

By the end of the evaluation, we had responses from 55 participants which we analysed from four different angles:

1. The configurations of the VC used in the evaluation
2. The different venue layouts used to test the performance of the VC's framing choices
3. The experience of the participants based on their academic background (year of study)
4. The experience of participants based on their average weekly lecture viewing time

With each of these angles of analysis, we compared how the proportions of agreement changed across the categories to determine any patterns in the data.

We found the VC satisfied the three major heuristics regardless of the VC configuration type, but 'Question 1', 'Question 3', and 'Question 8' were not significant for this angle of analysis. When looking at the remaining questions, we notice that all of them have a negative reception. 'Configuration 1' was set up such that the VC reacts more easily to changes in the presenter's movements which caused the VC to transition more often. This was done to test how participants would receive this behaviour and compare it to the 'Control' and 'Configuration 2'. From the results, we see that this change in the VC's configuration settings had no discernable difference to the way participants responded to the questions since the proportions between 'Configuration 1' and 'Configuration 2' are roughly the same across Questions 4, 5, 6, and 7.

When analysing the results from the angle of Venue Layout we found that, while some venues were more positively received than others, the VC satisfied the first three questions (i.e. the major heuristics) in all four venue layouts. 'Venue 2' performed the best of the four venue layouts despite the students moving in the audience and the content on the projector screens being small. The presenter paced around a lot in this venue, which caused the VC to change its framing more frequently. It could be that the framing changes (resulting from the presenter's pacing) are what caused participants to receive Venue 2 videos less negatively than the others.

'Venue 4' performed the worst of the four venue layouts because of the size of the venue and the presenter's lack of movement. When participants were shown videos from Venue 4 using the 'Control' they were unable to see the presenter's facial expressions and there was no board use. The presenter's movements were always close to the podium, which caused the VC to zoom in and out repeatedly as the presenter paced, affecting how participants responded. We observe that the more positively the first three questions are answered, the more negatively the remaining questions are answered for this angle of analysis (and vice versa).

We analysed the data from the angle of participant background to determine how experience with watching lecture recordings influenced our findings. We found that more experienced participants (in terms of their year of study) were more critical of the VC than less experienced participants and we see fewer 'Strongly Agree' answers and more 'Strongly Disagree' responses. Even with this decrease in positive reception, the VC still received a positive reception for the main heuristics in all background categories. We noticed that the '3rd Year' category of participants had the highest proportions of positive responses in the minor heuristics (Questions 4 through 8).

We see that the '3rd Year' category is the balance between year of study and average weekly hours of lecture viewing time. The 'Staff' category has fewer participants than all the other categories, so we cannot make too many comparisons with it regarding the other categories. We can conclude that participants in the '3rd Year' category have the most experience with watching lecture videos and they gave the most positive responses to the VC's output videos.

In terms of the average weekly time spent watching lecture videos, we found that the responses to Questions 1, 3, and 8 were not significant. We conclude that the VC satisfied the major heuristic of guiding the viewer's attention to what is important as we see that Question 2 received positive responses in all four angles of analysis. While Questions 1, and 3 are not significant, we observe a similarly positive reception as in Question 2 but we cannot conclude that these heuristics were successful. For the remaining heuristics, we conclude that the VC did not satisfy the majority of participants regardless of their average weekly viewing time. More work is required to improve the VC in these aspects.

6.1 Limitations

After analysing the results we notice that, while the VC succeeded in meeting its three major heuristics, it failed to satisfy participants in the minor heuristics.

This means that the VC is meeting the minimum requirements for success but it is not performing well in the more nuanced aspects of its performance. One of the leading causes for this is the variations in venue layouts which makes it difficult to generalise the VC's functionality such that it satisfies the minor heuristics. Other limitations include the number of participants, the number of video clip pairs shown to participants, the lack of open-ended questions in the questionnaire, and there is room for improvement in the design of the evaluation.

While it is useful to look at participant experience to determine if the feedback can be separated according to their backgrounds and average weekly lecture viewing times, it is not appropriate to mention this as a limitation to the study since the VC's objective is to satisfy its heuristics to as general an audience as possible.

6.2 Future Work

Question 4 in the evaluation tested participants' overall impression of the VC. We found that participants were not satisfied with the VC's camera work, but the questions do not indicate why. Future work must, therefore, include open-ended questions to allow participants to provide their perspectives on the VC's performance and how it could be improved.

Future work on this research could vary the VC configurations to include settings where the board content is prioritised over the presenter. Results from this configuration could determine if participants were more interested in the content being written on the board or if they wanted to see the presenter all the time.

We observe a negative reception of Configuration 1 in the VC's evaluation. While we suspect the increased camera motion from the close-up framing to be the cause of this result, more work is required to confirm this.

Future iterations of this research should exclude irrelevant questions from the matrix on a 'per video pair' basis, thereby restricting negative responses to genuine disagreement with the assertion of the question.

Further investigation is required to determine how closely our four angles of analysis are linked to one another such that we can better understand how participant experience affects the reception of the VC's output.

Our analysis for Question 6 indicates that participants were unsatisfied with the 'zoom and centre' control of the VC's framing choices. More work is required to determine if the zooming, centre framing, or both aspects are responsible for this negative reception.

6.3 Closing Summary

We set out to create a VC that can guide the viewer's attention to what is important while limiting the frequency of framing transitions to happen only when it is necessary and moving with smooth acceleration and deceleration when a framing transition does occur. These were the major heuristics which the VC used to inform its framing choices and we included some minor heuristics which were more difficult to implement and test due to the variety of lecture venue layouts and the complexity of evaluating these behaviours.

We found that the VC succeeded in satisfying the majority of evaluation participants with the major heuristics regardless of the VC's configuration, venue layout, participant background, or average weekly lecture viewing time. The VC failed to satisfy the majority of participants with the minor heuristics regardless of the VC's configuration, venue layout, participant background, or average weekly lecture viewing time.

While the VC has met the minimum requirements by succeeding with the major heuristics, more work is required to improve (and better test) the remaining minor heuristics.

Appendix A

Additional resources and explanations

A.1 More on the use of the dot product

This section elaborates on how we determine approximate collinearity (where lines have similar gradients) using the dot product.

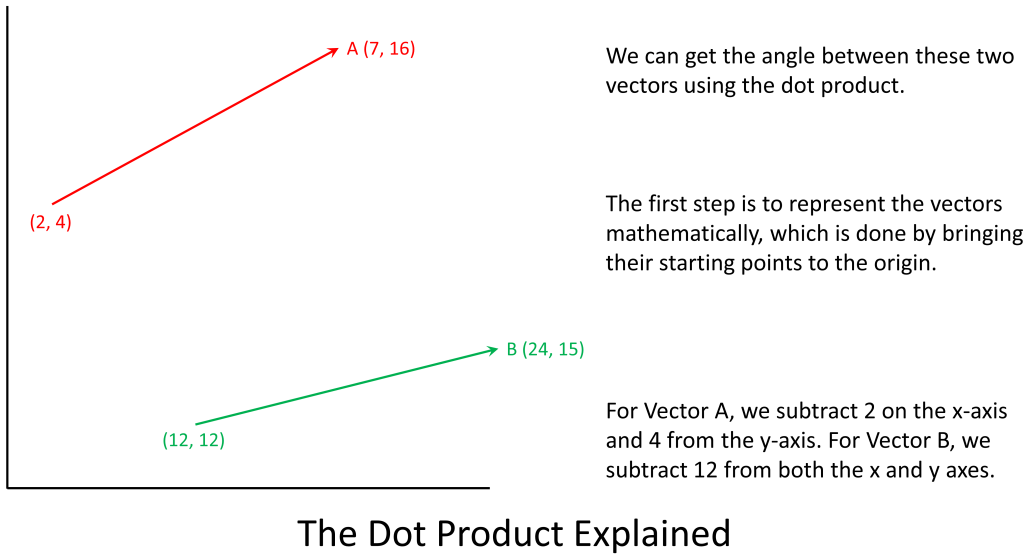


FIGURE A.1: Two example vectors at their original coordinates

Consider the example in Figure A.1 Where two vectors are shown. These vectors represent the lines of best fit drawn for two neighbouring time intervals in the usage information of a board in the venue.

To make it easier to know the angle between these vectors, we must translate them both to the origin.

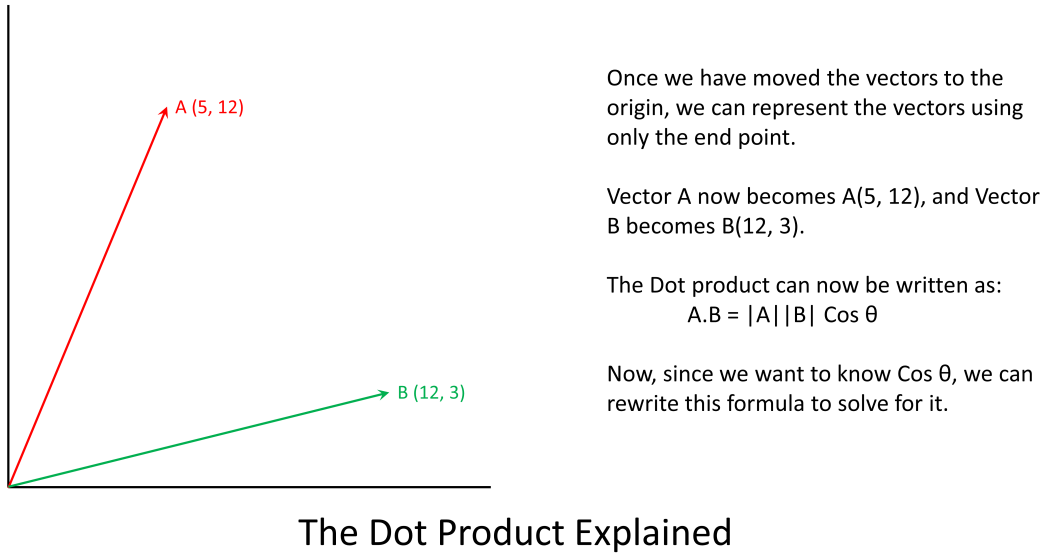


FIGURE A.2: Example vectors translated to the origin

After translating the vectors such that their starting points coincide with the origin, we can calculate their dot product.

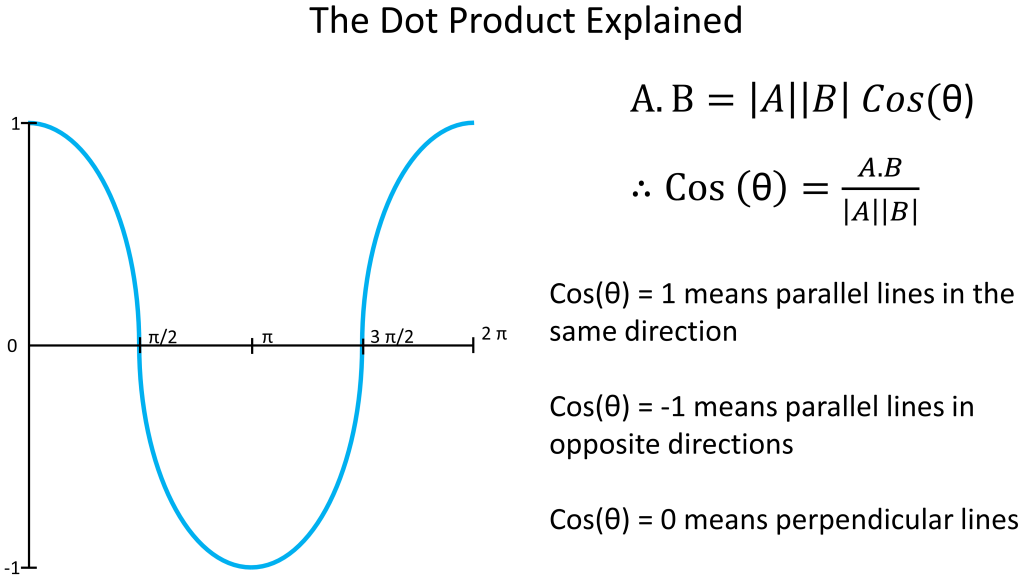


FIGURE A.3: Solve for $Cos\theta$ in the dot product formula

We re-organise the formula for the dot product such that we solve for the Cosine of the angle between the vectors. In Figure A.3 we show how the answer we get from solving for $Cos\theta$ in the equation can be interpreted to inform us about the collinearity of the vectors being compared. In the equation, we resolve $A.B$ using the alternative formula A.1 where we add the products of the corresponding vector coefficients:

$$A.B = (a_1)(b_1) + (a_2)(b_2) + ... + (a_n)(b_n) \tag{A.1}$$

In our case, we are not concerned about the direction of the vectors, since all the line segments are moving forward in time, which means we can take the absolute value of the answer to indicate whether the lines being compared are approximately collinear.

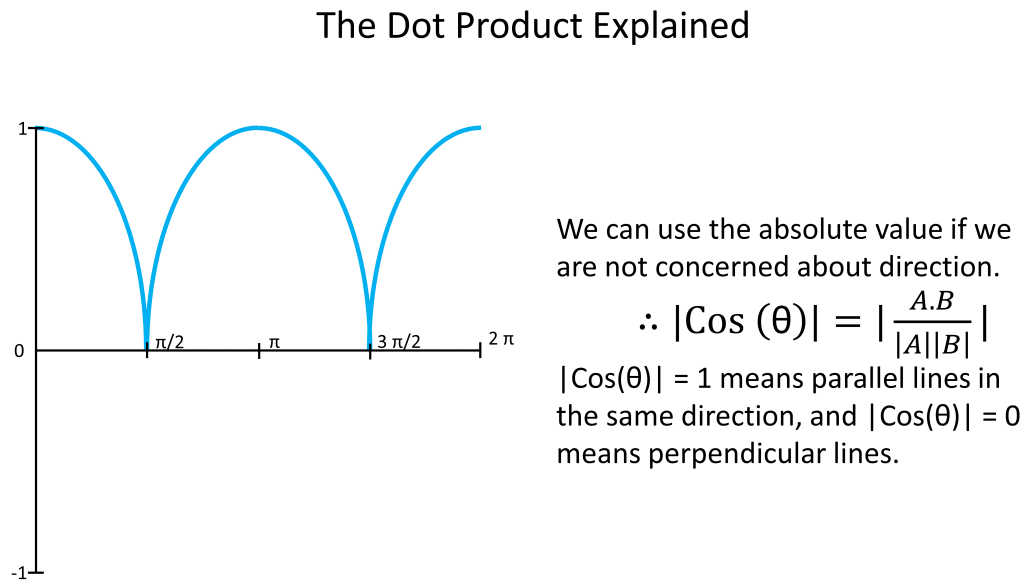


FIGURE A.4: Absolute Value of *Cosθ* removes negative results

Values closer to one indicate nearly full collinearity (where one is complete collinearity) and values closer to zero indicate almost no collinearity (where zero is complete perpendicularity). Collinear vectors are both parallel and in the same direction but, since we are looking at the changes in feature count over time, all the vectors are within the same π radians of rotation, namely all positive values of x, therefore we are not concerned about directionality, and all the vectors can be treated as line segments. In the case of values close to zero, we must keep these neighbouring line segments separate, and in the case of values close to one, we must join them.

A.2 Important classes of the Virtual Cinematographer

In this section, we include a detailed breakdown of the most important classes in the VC.

The PresenterAtFrame Class			
Variables		Methods	
Data Type	Variable Name	Return Type	Method Name
bool	deleted	Contructor	PresenterAtFrame
bool	gesturing	Constructor	PresenterAtFrame
bool	gesturingLeft	Copy Constructor	PresenterAtFrame
bool	gesturingRight	void	draw
int	PresenterID		
int	PartnerIndex		
int	circleRadius		
float	weight		
Vec41	gesturingLineLeft		
Vec4i	gesturingLineRight		
pair<CV::Point, int>	circleParameters		
CV::Point	circleCentre		
CV::Rect_<double>	presBoundingBox		

TABLE A.1: The variables and methods of the PresenterAtFrame Class

In table A.1 we see that there are still variables and data structures set up to handle gesture information. This functionality was not removed and was not used in the evaluation of the VC due to the limitations of temporal difference tracking when it comes to detecting gestures. Future work could improve on the implementation of gesture management in the VC provided another means of gesture detection is employed.

The Board Class			
Variables		Methods	
Data Type	Variable Name	Return Type	Method Name
bool	updated	explicit Contructor	Board
int	baordIndex	copy Constructor	Board
int	partnerID	Constructor	Board
int	numFeatures	void	addBoardLink
double	usageConfidence	void	consolidateLink
pair<int, int>	newBoardIndex	void	draw
CV::Scalar	colour		
CV::Rect_<double>	board		
Vector<BoardLink>	links		

TABLE A.2: The variables and methods of the Board Class

The ContextFrame Class			
Variables		Methods	
Data Type	Variable Name	Return Type	Method Name
int	chunkID	explicit Constructor	ContextFrame
int	frameNumberA	explicit Constructor	ContextFrame
int	frameNumberB	Constructor	ContextFrame
bool	multiPresenter	copy Constructor	ContextFrame
bool	containsBoardData	void	setEnclosingContext
bool	usageRelevant	void	setDebugEnclosingContext
bool	hasActiveBoards	void	draw
Rect_<double>	enclosingContext	void	consolidateAllLinks
Rect_<double>	initialChunk	bool	allPresentersDeleted
Rect_<double>	averageActiveRegion	Operator Overload	operator<
CV::Rect_<double>	getLastRealTimeFrame		
Vector<PresenterAtFrame>	presenters		
Vector<Board>	boardsInFrame		
Vector<Board>	enclosingContextExpansionVector		
Vector<PresenterAndBoardLink>	presenterAndBoardLinks		
Vector<int>	chosenBoardID		
Vector<int>	chosenPresentersIDs		
Vector<CV::Rect_<double>>	realtimeFrames		
Vector<pair<String, Rect_<double>>>	contextModificationHistory		

TABLE A.3: The variables and methods of the ContextFrame Class

The ContextFrame Class			
Variables		Methods	
Data Type	Variable Name	Return Type	Method Name
class	Chunk	Explicit constructor	ContextFramesManager
class	FittedLine	void	unwrapRegionRectangles
class	Usage	void	populatePresence
class	Dot	void	displayPresence
class	ActiveBoard	void	initialise
bool	favouringPresenter	void	binBoards
int	inputNumberOfFrames	void	addContextFrame
int	totalFramesToStore	void	setAllBoardIndices
int	skipFrames	void	setBoardIDs
int	inputVideoWidth	bool	checkAllRealTimeFramesLimit
int	inputVideoHeight	unsinged long	getSize
int	outputVideoWidth	Vector<CV::Rect_<double>>	unwrapRealtimeFrames
int	outputVideoHeight	Vector<vector<Board>>	getAllBoards
double	biasWeighting		
double	transitionLazinessFactor		
double	inputFPS		
String	JSON_FileName		
String	baseInputFileName		
String	debugOutVideoDir		
String	cropFilesOutDir		
String	videoFolder		
String	inputVideoFileName		
Vector<ContextFrame>	contextFrames		
Vector<float>	enclosingContextHeights		
Vector<double>	averageContextHeights		
Vector<double>	maximumContextHeights		
Vector<Chunk>	chunks		
Vector<ActiveBoards>	activeBoards		
Vector<pair<CV::Rect_<double>, bool>>	outputFrames		
Rect_<double>	lastRealtimeFrame		
Rect	gloablCrop		
Size	coordinateSpace		
map<int, int>	presenterPresence		

TABLE A.4: The variables and methods of the ContextFramesManager Class

A.3 Detailed analysis of the evaluation results tables

This section provides a more detailed analysis of the percentages and proportions of agreement in the VC's user evaluation. The results from these tables are condensed into stacked bar graphs in Chapter 5.

The colour scheme used in the tables of this section (detailed below) and is used to highlight the positive, negative, and neutral responses according to the percentage of responses.

- Bright green indicates percentages greater than 20% in the 'Strongly Agree' and 'Agree' categories.
- Pale green indicates percentages between 10% and 20% in the 'Strongly Agree' and 'Agree' categories.
- Bright yellow indicates percentages greater than 20% in the 'Neutral' category.
- Pale yellow indicates percentages between 10% and 20% in the 'Neutral' category.
- Bright red indicates percentages greater than 20% in the 'Strongly Disagree' and 'Disagree' categories.
- Pale red indicates percentages between 10% and 20% in the 'Strongly Disagree' and 'Disagree' categories.

A.3.1 Results pertaining to the VC's Configurations

The following table shows how each of the VC's Configurations was received by the participants in the evaluation.

Question Number		1	2	3	4	5	6	7	8
Control	Strongly Agree	42.9	33.8	51.3	0.0	2.7	0.0	2.5	0.0
	Agree	34.3	63.1	23.1	8.5	10.8	16.3	30.0	6.7
	Neutral	20.0	1.5	17.9	25.5	13.5	16.3	0.0	0.0
	Disagree	2.9	1.5	7.7	4.3	18.9	9.3	7.5	3.3
	Strongly Disagree	0.0	0.0	0.0	61.7	54.1	58.1	60.0	90.0
Config 1	Strongly Agree	31.5	58.0	27.1	0.8	0.0	0.8	0.5	1.1
	Agree	48.7	31.6	46.7	13.6	16.2	12.9	16.3	8.5
	Neutral	9.6	0.0	14.8	2.7	26.1	0.4	4.5	3.7
	Disagree	9.0	10.1	8.7	12.1	12.6	15.4	8.4	2.7
	Strongly Disagree	1.2	0.3	2.7	70.8	45.1	70.4	70.3	84.0
Config 2	Strongly Agree	35.3	52.8	29.1	0.7	0.0	1.8	0.8	0.3
	Agree	46.4	36.6	42.0	15.7	12.2	16.0	14.0	8.9
	Neutral	9.4	0.7	16.6	6.7	19.5	2.1	1.0	1.1
	Disagree	6.6	9.8	9.4	13.0	18.2	19.5	13.7	6.4
	Strongly Disagree	2.3	0.2	2.9	63.9	50.1	60.6	70.5	83.3

TABLE A.5: User reception (percentage) related to the VC's configuration type

When looking at the results from Table A.5, we must note that the findings for Questions 1, 3, and 8 are not significant (as shown in table 5.1) so we can only make conclusions from the remaining questions' cross-tabulations.

We observe a similar pattern with the responses to Question 3. There is a higher level of agreement in the 'Control' configuration ('Strongly Agree') than there is in either of the VC's configurations ('Agree').

We observe the level of agreement for Question 8 is consistent across all configurations ('Strongly Disagree') and that more than 80% of the responses are 'Strongly Disagree'. The 'Control' configuration had the highest proportion (90%) of responses in the 'Strongly Disagree' category.

We see a mostly positive response for all three configurations when looking at Question 2, where 42.9% of participants answered 'Strongly Agree' and 34.3% answered 'Agree' for the 'Control' configuration. For 'Configuration 1', we see that 31.5% of participants answered 'Strongly Agree' and 48.7% answered 'Agree'. 35.3% of participants answered 'Strongly Agree' and 46.6% answered 'Agree' for 'Configuration 2'.

We see that Questions 4, 5, 6, 7, and 8 have more negative responses than positives, and the findings in Question 8 are not significant. This means we should only make conclusions based on the findings from the cross-tabulations of Questions 4, 5, 6, and 7.

The results for Question 4 show that participants responded negatively to the 'Control' configuration with 61.7% responding with 'Strongly Disagree' and 4.3% responding with 'Disagree', while 70.8% responded 'Strongly Disagree' and 12.1% responded 'Disagree' in 'Configuration 1', and 63.9% responded 'Strongly Disagree' and 13.0% responded 'Disagree' for 'Configuration 2'.

The results for Question 5 show that 54.1% of participants answered 'Strongly Disagree' and 18.9% answered 'Disagree' for the 'Control' configuration. With 'Configuration 1' we see that 45.1% responded 'Strongly Disagree' and 12.6% responded 'Disagree' for 'Configuration 1'. 50.1% of participants responded 'Strongly Disagree' for 'Configuration 2' and 18.2% responded 'Disagree'.

When looking at Question 6, we see that 58.1% of participants responded 'Strongly Disagree' and 9.3% responded 'Disagree' to the 'Control' videos. For 'Configuration 1' 70.4% of participants responded 'Strongly Disagree' and 15.4% responded 'Disagree'. 60.6% responded 'Strongly Disagree' for 'Configuration 2' and 19.5% responded 'Disagree'.

Finally, we look at the responses to Question 7. For the 'Control' videos, 60.0% of participants responded 'Strongly Disagree' and 7.5% responded 'Disagree'. Then, for 'Configuration 1', we see 70.3% of participants responded 'Strongly Disagree' and 8.4% responded 'Disagree'. 70.5% of participants responded 'Strongly Disagree' for 'Configuration 2' and 13.7% responded 'Disagree'.

A.3.2 Results pertaining to the venue types used in the study

Table A.6 shows how each of the lesson venues was received by the participants for each question of the evaluation.

In Question 1 we see the highest agreement coming from Venue 4, the second-highest levels of agreement come from Venue 3, followed by Venue 1, and the lowest agreement from Venue 2. In all 4 venues, we see low levels of disagreement regarding this question.

In Question 2 we see the highest level of agreement coming from the responses to the videos recorded in Venue 4 with Venue 3 having the second-highest levels of agreement, followed by Venue 1, and Venue 2 has the lowest levels of agreement for this question.

We also notice that almost all of the responses from Venues 3 and 4 for this question are either 'Strongly Disagree' or 'Agree' with very few responses in disagreement or neutrality. We also see that the proportion of 'Disagree' responses nearly match the proportion of 'Agree' responses for Venue 2.

Venue 4 has the highest levels of agreement for Question 3, with the second-highest levels of agreement coming from Venue 3, followed by Venue 1, and Venue 2 has the lowest levels of agreement for the question. We notice that the proportions of the responses to Venues 1 and 2 are more distributed among the 5 response categories compared to Venues 3 and 4. We also notice that, for each venue, the proportion of 'Neutral' responses is higher than the proportions of 'Strongly Disagree' and 'Disagree' responses.

In Question 4 we see that 64.9% of the responses from Venue 1 responded 'Strongly Disagree' and 13.9% responded 'Disagree'. 51.6% of participants responded 'Strongly Disagree' for Venue 2 and 23.3% responded 'Disagree'. For Venue 3, we see that 76.9% responded 'Strongly Disagree and 5.5% responded 'Disagree'. Finally, for Venue 4, we see that 71.5% of participants responded 'Strongly Disagree' and 3.6% responded 'Disagree'.

Question Number		1	2	3	4	5	6	7	8
Venue 1	Strongly Agree	33.6	51.6	29.6	0.5	0.0	1.0	1.2	0.7
	Agree	48.3	35.8	41.6	16.0	12.6	13.0	16.5	5.8
	Neutral	9.3	0.0	16.0	4.6	10.9	4.0	1.2	1.4
	Disagree	7.7	11.8	10.4	13.9	16.9	19.0	9.1	2.2
	Strongly Disagree	1.2	0.8	2.4	64.9	59.6	63.0	72.0	89.9
Venue 2	Strongly Agree	24.2	53.6	17.6	0.9	0.9	1.5	1.0	1.1
	Agree	40.1	24.1	39.3	19.1	13.1	18.8	23.0	14.4
	Neutral	13.5	1.8	19.7	5.1	20.6	1.5	0.5	1.1
	Disagree	15.9	20.5	17.2	23.3	30.4	30.7	18.3	12.8
	Strongly Disagree	6.3	0.0	6.3	51.6	35.0	47.5	57.1	70.6
Venue 3	Strongly Agree	37.4	57.4	33.3	0.5	0.0	1.1	1.3	0.7
	Agree	49.6	36.3	45.0	9.5	18.0	13.6	11.9	5.6
	Neutral	10.2	0.8	14.6	7.5	41.0	1.1	0.0	0.0
	Disagree	2.4	5.5	6.3	5.5	9.7	11.3	9.4	2.1
	Strongly Disagree	0.4	0.0	0.8	76.9	31.3	72.9	77.5	91.6
Venue 4	Strongly Agree	44.3	49.4	37.9	1.2	0.0	2.6	0.0	0.0
	Agree	46.7	48.3	44.0	14.2	8.3	13.2	9.9	9.5
	Neutral	6.1	0.0	14.0	9.1	2.5	3.3	8.4	5.6
	Disagree	2.9	2.2	2.5	3.6	3.3	6.6	8.4	0.0
	Strongly Disagree	0.0	0.0	1.6	71.5	86.0	74.3	73.3	84.9

TABLE A.6: User reception (percentage) related to the Lecture Venue Layout

There are higher levels of disagreement in Question 4 compared to the levels of agreement. We see that Venue 3 has the highest levels of disagreement and Venue 1 has the second-highest levels of disagreement. Venue 4 has the second-lowest levels of disagreement with Venue 2 having the lowest levels of disagreement of the 4 venues for Question 4. We notice that Venue 2 (while having the smallest proportion of 'Strongly Disagree' and 'Disagree' responses) also has the highest proportion of 'Agree' responses compared to the other venues.

In Question 5 we see that Venue 4 has the highest levels of disagreement with Venue 1 in second place, Venue 2 in third place, and Venue 3 with the lowest levels of disagreement among the venues for this question. We notice that Venue 3 has a high proportion of 'Neutral' responses in comparison to the other venues and it also has the highest levels of agreement of the venues for this question.

There is a strong negative response to Question 6 for each of the venues with Venue 3 having the highest levels of disagreement and Venue 1 with the second-highest. Venue 4 has the second-lowest levels of disagreement, and Venue 2 has the lowest levels of disagreement among the venues for this question. We observe Venue 2 has the lowest proportion of 'Strongly Disagree' responses of all 4 venues for this question.

For Question 7 we see that Venue 3 has the highest levels of disagreement of all 4 venues for this question with Venue 4 in second place. Venue 1 has the second-lowest levels of disagreement, and Venue 2 has the lowest levels of disagreement. We observe Venue 2 has the highest proportion of 'Agree' responses of the 4 venues for this question.

Question 8 received the largest proportion of negative responses per venue of all 8 questions. 89.9% of responses from Venue 1 were 'Strongly Disagree' and 2.2% were 'Disagree'. 70.6% of responses from Venue 2 were 'Strongly Disagree' and 12.8% were 'Disagree'. From the Venue 3 venue, we see that 91.6% of participants responded 'Strongly Disagree' and 2.1% responded 'Disagree'. Finally, from Venue 4, we see that 84.9% of responses were 'Strongly Disagree' and none of the responses was 'Disagree'.

Question 8 has high levels of disagreement for each venue. Venue 3 has the highest, followed by Venue 1, Venue 4, and Venue 2 has the lowest levels of disagreement of all the venues for this question. We notice that the response category with the largest proportions for each venue is the 'Strongly Disagree' response category.

A.3.3 Results pertaining to participant background

Question Number		1	2	3	4	5	6	7	8
1st Year	Strongly Agree	44.7	46.4	32.6	0.0	0.0	1.7	0.7	0.8
	Agree	35.4	46.0	40.0	18.0	11.5	12.9	19.2	9.8
	Neutral	9.7	0.8	16.5	6.6	21.8	1.7	3.4	3.8
	Disagree	6.3	6.3	8.7	10.9	17.8	21.3	11.0	6.8
	Strongly Disagree	3.8	0.4	2.2	64.5	48.9	62.4	65.8	78.8
2nd Year	Strongly Agree	43.4	47.5	49.7	0.0	0.0	0.8	0.0	0.0
	Agree	35.7	46.9	22.8	18.3	14.7	23.3	21.0	8.0
	Neutral	14.3	0.0	11.4	4.2	27.9	5.8	3.0	3.4
	Disagree	5.5	5.6	9.0	10.8	17.1	12.5	4.0	2.3
	Strongly Disagree	1.1	0.0	7.2	66.7	40.3	57.5	72.0	86.4
3rd Year	Strongly Agree	30.3	52.4	26.7	1.7	1.0	2.9	2.1	1.0
	Agree	49.0	35.9	47.3	16.4	12.4	19.0	18.6	14.6
	Neutral	11.6	0.7	16.0	0.0	24.8	1.9	1.0	0.0
	Disagree	7.7	11.0	8.0	14.7	19.0	15.2	18.6	7.3
	Strongly Disagree	1.3	0.0	2.0	67.2	42.9	61.0	59.8	77.1
Postgrad	Strongly Agree	29.4	56.8	21.0	0.4	0.4	1.7	1.0	0.6
	Agree	50.7	27.9	44.4	14.3	17.7	13.5	13.9	7.9
	Neutral	8.1	0.7	22.0	5.7	16.0	2.6	1.5	1.1
	Disagree	10.1	14.3	10.5	15.2	18.6	22.2	17.3	6.2
	Strongly Disagree	1.7	0.3	2.0	64.3	47.2	60.0	66.3	84.2
Staff	Strongly Agree	0.0	80.0	0.0	5.0	0.0	0.0	0.0	0.0
	Agree	95.0	5.0	84.2	5.0	5.3	10.0	10.0	5.0
	Neutral	0.0	0.0	0.0	0.0	57.9	0.0	10.0	5.0
	Disagree	0.0	15.0	15.8	15.0	15.8	5.0	5.0	0.0
	Strongly Disagree	0.0	0.0	0.0	75.0	21.1	85.0	75.0	90.0

TABLE A.7: User reception (percentage) related to participant background

Table A.7 shows how participants responded to each of the questions in the evaluation grouped according to participant background.

Question 1 was phrased as follows, 'The camera tracked the presenter smoothly'. It asked participants if the VC was tracking the presenter smoothly. We notice a larger proportion of participants responded in agreement ('Strongly Agree' and 'Agree') across all participant background groups. The highest proportion of positive responses came from the staff member background category. This is because there were so few participants with this background in the evaluation. The second-highest proportion is shared (both have 80.1% positive responses) by the first-year and postgrad background categories but the first-year participants had a larger portion of 'Strongly Agree' responses than the postgrad participants. The 3rd-year category had the second-lowest proportion and the 2nd-year students had the lowest. We notice that the percentages of 'Strongly Agree' responses decrease as the participant background changes from 1st-year through to staff.

Question 2 was phrased, 'I could see what I wanted to watch.' We see that there are high levels of agreement on this question across all participant backgrounds. The 2nd-year category has the highest percentage of positive responses, followed by the 1st-year category, 3rd-year, staff, and the postgrad category had the lowest percentage of positive responses. For this question, we notice that the percentage of 'Strongly Agree' responses increases as the categories change from 1st-year to staff which is the inverse of what we observe in Question 1.

Question 3 was phrased, 'I like the frequency with which the camera shots changed.' and we notice that there is a high percentage of positive responses for this question across all the background categories. The highest percentage comes from the staff category, followed by the 3rd-year, 1st-year, 2nd-year, and the postgrad category had the lowest percentage of positive responses. The postgrad category has a large percentage of neutral responses (22%) which brings the positive response percentages down.
It also has the most spread-out set of responses among all 5 categories. We observe that the

percentages of positive responses across the 5 categories do not change in a linear fashion as with Questions 1 and 2.

Question 5 was phrased as follows, 'I was able to follow with what was written on the board.'. For this question, we see that there is a higher proportion of negative responses (responses of 'Strongly Disagree' and 'Disagree') and we note that the 1st-year category has the highest percentage of negative responses followed by postgrad, 3rd-year, 2nd-year, and the staff category had the lowest percentage of negative responses. We note that four of the five categories had large percentages of 'Neutral' responses (21.8% up to 57.9%), which could indicate that some participants were unsure about their response to this question.

Question 7 checked if participants could see the presenter's facial expressions. We note that this question (as with Question 5 above) has a large percentage of negative responses and the postgrad category has the highest percentage, followed by staff, 3rd-year, 1st-year, and the 2nd-year category has the lowest percentage of negative responses. We note that all categories have high percentages of 'Strongly Disagree' which means that most of the participants were not able to see the presenter's facial expressions. This could be due to the VC's having to zoom out to accommodate more of the lecture environment. Zooming out this much then reduces the size of the presenter's face and, consequently, makes it more difficult to see the facial expressions.

A.3.4 Results pertaining to participant weekly viewing time

Question Number		1	2	3	4	5	6	7	8
<2 hours	Strongly Agree	41.0	44.9	37.8	0.8	0.0	0.6	0.4	0.4
	Agree	39.8	41.8	31.6	16.2	16.4	17.6	14.8	9.3
	Neutral	6.6	0.6	17.8	11.5	22.2	2.8	2.5	1.6
	Disagree	9.7	12.3	10.1	15.1	20.4	21.7	16.2	6.5
	Strongly Disagree	2.9	0.4	2.6	56.3	41.0	57.2	66.2	82.2
2 hours	Strongly Agree	19.7	75.7	12.2	0.0	0.0	0.7	0.8	0.8
	Agree	60.6	20.7	60.5	11.9	10.7	14.8	21.1	8.5
	Neutral	14.8	0.0	10.9	3.0	23.1	2.2	1.6	1.7
	Disagree	4.9	3.6	8.8	8.1	8.3	11.1	5.7	4.2
	Strongly Disagree	0.0	0.0	7.5	77.0	57.9	71.1	70.7	84.7
3 hours	Strongly Agree	25.0	63.1	19.6	0.8	0.8	1.6	2.6	2.0
	Agree	60.5	29.5	61.4	16.2	10.5	6.3	12.2	10.8
	Neutral	9.2	0.7	14.4	11.5	20.2	2.4	1.7	0.0
	Disagree	5.3	6.7	3.9	15.1	12.9	15.1	4.3	0.0
	Strongly Disagree	0.0	0.0	0.7	56.3	55.6	74.6	79.1	87.3
>3 hours	Strongly Agree	37.5	49.5	31.3	0.0	0.6	3.9	0.8	0.0
	Agree	41.1	40.7	39.6	18.0	12.4	15.8	16.9	8.3
	Neutral	13.8	0.9	17.1	1.2	18.6	2.0	2.4	4.1
	Disagree	4.9	8.9	10.6	14.3	17.4	17.5	14.5	6.6
	Strongly Disagree	2.7	0.0	1.4	66.5	50.9	60.5	65.3	81.0

TABLE A.8: User reception (percentage) related to participant viewing time

Question 1 tested if the VC tracked the presenter smoothly. We note that there is a higher percentage of positive responses ('Strongly Agree' and 'Agree') for this question across all the 'viewing time' categories. We also see that, for this question, the '3 Hours' category had the highest percentage of positive responses followed by 'Less Than 2 Hours', '2 Hours', and the 'More than 3 hours' had the lowest percentage of the 4 categories. The difference between the percentages of the '3 Hours' and 'More Than 3 hours' categories is 6.9%. We observe the strongest agreement from the 'Less Than 2 Hours' and 'More Than 3 Hours' categories, which were more balanced in the levels of agreement with a slight bias towards 'Agree' rather than 'Strongly Agree'. This bias is even more exaggerated in the '2 Hours' and '3 Hours' categories.

Question 2 tested if participants were able to see what they wanted to watch when viewing the videos in the evaluation. We observe that there is a higher percentage of positive responses for this question across all the 'viewing time' categories. We note that the '2 Hours' category had the highest percentage of positive responses followed by '3 Hours', 'More Than 3 Hours', and 'Less Than 2 Hours' had the lowest percentage of positive responses. The difference between the percentages of the '2 Hours' and 'Less Than 2 hours' categories is 9.7%. We see that the 'Less Than 2 Hours' and 'More Than 3 Hours' categories have the most balanced distribution between 'Strongly Agree' and 'Agree' responses with a slight bias towards 'Strongly Agree' rather than 'Agree'. This bias was exaggerated for the '2 Hours' and '3 Hours' categories.

Question 3 asked participants if they liked the frequency of camera transitions in the videos during the evaluation. We notice that there is a higher percentage of positive responses from all 4 categories for this question. We observe that the '3 Hours' category had the highest percentage of positive responses followed by '2 Hours', 'More Than 3 Hours', and the 'Less Than 2 Hours' category had the lowest percentage of positive responses. The difference between the percentages of the '3 Hours' and 'Less Than 2 hours' categories is 1.5%. The 'Less Than 2 Hours' and 'More Than 3 Hours' categories have the most balanced distribution between 'Strongly Agree' and 'Agree' responses. We see a strong bias for the '2 Hours' and '3 Hours' categories towards 'Agree' rather than 'Strongly Agree'.

Question 4 tested the overall impressions the VC gave participants about its control of the camera and framing decisions. For this question, we see that there is a higher percentage of negative responses ('Strongly Disagree' and 'Disagree') from all 4 of the 'viewing time' categories. We notice that the '2 Hours' category had the highest percentage of negative responses followed by 'More Than 3 Hours', and the 'Less Than 2 Hours' category had the same percentage of negative responses as '3 Hours'. The difference between the percentages of the '2 Hours' and 'Less Than 2 Hours' categories is 13.7%. There is a strong bias towards the 'Strongly Disagree' response rather than 'Disagree' for all 4 categories. We also see that the percentage of 'Agree' responses is higher than the percentage of 'Disagree' responses for all categories in this Question.

Question 5 tested if participants could follow what was written on the boards. For this question, we notice that there is a higher percentage of negative responses for all 4 'viewing time' categories. We observe that the '3 Hours' category had the highest percentage of negative responses followed by 'More Than 3 Hours', '2 Hours', and the 'Less Than 2 Hours' category had the lowest percentage of negative responses. The difference between the percentages of the '3 Hours' and 'Less Than 2 Hours' categories is 7.1%. There is a strong bias towards 'Strongly Disagree' responses rather than 'Disagree' responses and we also note that all categories have a notable percentage of 'Neutral' responses.

Question 6 was phrased, 'The camera view was zoomed and centred appropriately'. We see that, for this question, there is a higher percentage of negative responses for all 4 categories. We note that the '3 Hours' category had the highest percentage of negative responses followed by '2 Hours', 'Less Than 2 Hours', and the 'More Than 3 Hours' category had the lowest percentage of negative responses. The difference between the percentages of the '3 Hours' and 'More Than 3 Hours' categories is 18.9%. There is a strong bias towards 'Strongly Disagree' responses rather than 'Disagree' responses.

Question 7 asked participants if they could see the presenter's facial expressions when watching the videos in the evaluation. For this question, there is a higher percentage of negative responses for all 4 categories. We observe that the '3 Hours' category had the highest percentage of negative responses followed by 'Less Than 2 Hours', 'More Than 3 Hours', and the '2 Hours' category had the lowest percentage of negative responses. The difference between the percentages of the '3 Hours' and '2 Hours' categories is 1%. There is a strong bias towards 'Strongly Disagree' responses rather than 'Disagree' responses. We also see that there is a notable percentage of 'Agree' responses for this question where the percentage of 'Agree' responses is higher than the percentage of 'Disagree' responses except for the 'Less Than 2 Hours' category.

Bibliography

[1] W. A. Wittich. "Wisconsin Physics Film Evaluation Project". In: *The American Mathematical Monthly* 66.5 (May 1959), p. 417. ISSN: 00029890. DOI: 10.2307/2308761. URL: https://www.jstor.org/stable/2308761?origin=crossref.

[2] Charles P Benner and Curtis A Rogers. "A new plan for instructing large classes in mathematics by television and films". In: *The Mathematics Teacher* 53.5 (1960), pp. 371–375. DOI: 10.2307/27956175. URL: http://www.jstor.org/stable/27956175.

[3] Frank R. Hartman. *Filmed lectures as supplementary and refresher education*. Tech. rep. 4. 1960, pp. 214–219. DOI: 10.1007/BF02713445.

[4] D. Goodhue. "Tape-recorded Lectures with Slide Synchronization A Description of the Method". In: *Journal of Biological Education* 3.4 (Dec. 1969), pp. 311–319. ISSN: 21576009. DOI: 10.1080/00219266.1969.9653601.

[5] Jock D Mackinlay, Stuart K Card, and George G Robertson. "Rapid controlled movement through a virtual 3D workspace". In: *Computer Graphics (ACM)* 24.4 (1990), pp. 171–176. ISSN: 00978930. DOI: 10.1145/97880.97898.

[6] Steven D. (Steven Douglas) Katz. *Film directing shot by shot : visualizing from concept to screen*. Los Angeles, USA: Michael Wiese Productions, 1991, pp. 1–474. ISBN: 0941188108.

[7] Peter Karp and Steven Feiner. "Automated presentation planning of animation using task decomposition with heuristic reasoning". In: *Proceedings - Graphics Interface* (1993), pp. 118–127. ISSN: 07135424.

[8] Desmond Keegan. *Open Universities-Their Rationale, Characteristics, and Prospects*. Tech. rep. 1994, pp. 18–32.

[9] F. Bodendorf, R. Grebner, and C. Langenbach. "Virtual lecture room - practice and experiences". In: *IEEE International Conference on Multi-Media Engineering Education - Proceedings*. IEEE, 1996, pp. 277–282. DOI: 10.1109/mmee.1996.570273.

[10] Stefan Leue. "Specifying Real-Time Requirements for SDL Specifications — A Temporal Logic-Based Approach". In: 1996, pp. 19–34. DOI: 10.1007/978-0-387-34892-6{_}2.

[11] Doug A. Bowman, David Koller, and Larry F. Hodges. "Travel in immersive virtual environments: an evaluation of viewpoint motion control techniques". In: *Proceedings IEEE 1997 Annual International Symposium on Virtual Reality*. Albuquerque, NM, USA, IEEE, 1997, pp. 45–52. ISBN: 0-8186-7843-7. DOI: 10.1109/vrais.1997.583043.

[12] Herng Yow Chen, Gin Yin Chen, and Jen Shin Hong. "Design of a web-based synchronized multimedia lecture system for distance education". In: *International Conference on Multimedia Computing and Systems -Proceedings* 2 (1999), pp. 887–891. DOI: 10.1109/mmcs.1999.778605.

[13] Sugata Mukhopadhyay and Brian Smith. "Passive Capture and Structuring of Lectures". In: *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*. MULTIMEDIA '99. Orlando, Florida, USA: ACM, 1999, pp. 477–487. DOI: 10.1145/319463.319690. URL: https://doi.org/10.1145/319463.319690.

[14] Michael Gleicher and J. Masanz. "Towards virtual videography". In: *Proceedings of the ACM International Multimedia Conference and Exhibition* (2000), pp. 375–378.

[15] Michael Gleicher and James Masanz. "Towards Virtual Videography (Poster Session)". In: *Proceedings of the eighth ACM international conference on Multimedia*. MULTIMEDIA '00. Marina del Rey, California, USA: ACM, 2000, pp. 375–378. ISBN: 1581131984. DOI: 10.1145/354384.354537.

[16] Qiong Liu, Yong Rui, Anoop Gupta, et al. "Automating Camera Management for Lecture Room Environments". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '01. Seattle, Washington, USA: ACM, 2001, pp. 442–449. DOI: 10.1145/365024.365310.

[17] Yong Rui, Liwei He, Anoop Gupta, et al. "Building an intelligent camera management system". In: *Proceedings of the ACM International Multimedia Conference and Exhibition*. MULTIMEDIA '01. Ottawa, Canada: ACM, 2001, pp. 2–11. ISBN: 1581133944. DOI: 10.1145/500144.500145.

[18] Michael Gleicher and Nicola Ferrier. "Evaluating Video-Based Motion Capture". In: *Proceedings of Computer Animation 2002 (CA 2002)*. Geneva, Switzerland: IEEE, 2002, pp. 70–80. URL: https://graphics.cs.wisc.edu/Papers/2002/GF02/videomocap.pdf.

[19] Michael L. Gleicher, Rachel M. Heck, and Michael N. Wallick. "A Framework for Virtual Videography". In: *Proceedings of the 2nd International Symposium on Smart Graphics*. Vol. 22. SMARTGRAPH '02. Hawthorne, New York, USA: ACM, 2002, pp. 9–16. ISBN: 1581135556. DOI: 10.1145/569005.569007. URL: http://dl.acm.org/citation.cfm?id=569005.569007.

[20] Anurag Mittal and Larry S. Davis. "M2 tracker: A multi-view approach to segmenting and tracking people in a cluttered scene". In: *International Journal of Computer Vision* 51.3 (2003), pp. 189–203. ISSN: 09205691. DOI: 10.1023/A:1021849801764.

[21] Yong Rui, Anoop Gupta, and Jonathan Grudin. "Videography for Telepresentations". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '03. Ft. Lauderdale, Florida, USA: ACM, 2003, pp. 457–464. ISBN: 1581136307. DOI: 10.1145/642611.642691. URL: http://portal.acm.org/citation.cfm?doid=642611.642691.

[22] Andrew Wallace. "The Calibration and Optimisation of Speed-Dependent Automatic Zooming". In: November (2003).

[23] Yong Rui, Anoop Gupta, Jonathan Grudin, et al. "Automating lecture capture and broadcast: Technology and videography". In: *Multimedia Systems* 10.1 (2004), pp. 3–15. ISSN: 09424962. DOI: 10.1007/s00530-004-0132-9. URL: https://link-springer-com.ezproxy.uct.ac.za/article/10.1007/s00530-004-0132-9.

[24] Sascha Konrad and Betty H.C. Cheng. "Real-time specification patterns". In: *Proceedings - 27th International Conference on Software Engineering, ICSE05*. 2005, pp. 372–381. DOI: 10.1145/1062455.1062526.

[25] Takao Yokoi and Hironobu Fujiyoshi. "Virtual camerawork for generating lecture video from high resolution images". In: *2005 IEEE International Conference on Multimedia and Expo*. Amsterdam, Netherlands: IEEE, 2005. ISBN: 0780393325. DOI: 10.1109/ICME.2005.1521532.

[26] Cha Zhang, Yong Rui, Li Wei He, et al. "Hybrid speaker tracking in an automated lecture room". In: *IEEE International Conference on Multimedia and Expo, ICME 2005*. Vol. 2005. Amsterdam: IEEE, 2005, pp. 81–84. ISBN: 0780393325. DOI: 10.1109/ICME.2005.1521365.

[27] Zhenqiu Zhang, Gerasimos Potamianos, Andrew Senior, et al. "A joint system for person tracking and face detection". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 3766 LNCS.506909 (2005), pp. 47–59. ISSN: 03029743. DOI: 10.1007/11573425{_}5.

[28] Andrea Cavallaro, Ruchira Chandrasekera, and Murtaza Taj. "Hands-On Experience in Image Processing: The Automated Lecture Cameraman". In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*. Vol. 3. ICASSP '07. Honolulu, HI, USA: IEEE, 2007, pp. 721–724. ISBN: 1-4244-0728-1. DOI: 10.1109/ICASSP.2007.366781. URL: http://ieeexplore.ieee.org/document/4217811/.

[29] Rachel Heck, Michael Wallick, and Michael Gleicher. "Virtual Videography". In: *ACM Transactions on Multimedia Computing, Communications, and Applications* 3.1 (2007), 4–es. ISSN: 15516857. DOI: 10.1145/1198302.1198306. URL: http://portal.acm.org/citation.cfm?doid=1198302.1198306.

[30] Fleming Lampi, Stephan Kopf, Manuel Benz, et al. "An Automatic Cameraman in a Lecture Recording System". In: *Proceedings of the International Workshop on Educational Multimedia and Multimedia Education*. Emme '07. Augsburg, Bavaria, Germany: ACM, 2007, pp. 11–18. ISBN: 9781595937834. DOI: 10.1145/1290144.1290148. URL: http://portal.acm.org/citation.cfm?doid=1290144.1290148.

[31] Marc Christie, Patrick Olivier, and Jean-Marie Normand. "Camera Control in Computer Graphics". In: *Computer Graphics Forum* 27.8 (Dec. 2008), pp. 2197–2218. ISSN: 01677055.

DOI: 10.1111/j.1467-8659.2008.01181.x. URL: http://doi.wiley.com/10.1111/j.1467-8659.2008.01181.x.

[32] Maree Gosper, David Green, Margot Mcneill, et al. *The Impact of Web-Based Lecture Technologies on Current and Future Practices in Learning and Teaching*. Tech. rep. 2008. URL: www.altc.edu.au..

[33] Fleming Lampi, Stephan Kopf, and Wolfgang Effelsberg. "Automatic Lecture Recording". In: *Proceedings of the 16th ACM International Conference on Multimedia*. MM '08. Vancouver, British Columbia, Canada: ACM, 2008, pp. 1103–1105. ISBN: 9781605583037. DOI: 10.1145/1459359.1459583. URL: https://doi.org/10.1145/1459359.1459583.

[34] Cha Zhang, Yong Rui, Jim Crawford, et al. "An Automated End-to-End Lecture Capture and Broadcasting System". In: *ACM Transactions on Multimedia Computing, Communications, and Applications* 4.1 (2008), pp. 1–23. ISSN: 1551-6857. DOI: 10.1145/1324287.1324293. URL: http://portal.acm.org/citation.cfm?doid=1324287.1324293.

[35] Christopher Brooks, Kristofor Amundson, and Jim Greer. *Detecting Significant Events in Lecture Video using Supervised Machine Learning*. Tech. rep. University of Saskatchewan, 2009, pp. 483–490. DOI: 10.3233/978-1-60750-028-5-483. URL: http://www.replay.ethz.ch/.

[36] Takayuki Nagai. "Automated lecture recording system with AVCHD camcorder and microserver". In: *Proceedings of the 2009 ACM SIGUCCS Fall Conference*. SIGUCCS'09. St. Louis, Missouri, USA, 2009, pp. 47–54. ISBN: 9781605584775. DOI: 10.1145/1629501.1629512. URL: http://dl.acm.org/citation.cfm?id=1629512.

[37] Zolqernine Othman, Mohammed Rafiq, and Abdul Kadir. "Comparison of Canny and Sobel Edge Detection in MRI Images". In: (2009), pp. 133–136. URL: https://pdfs.semanticscholar.org/3d97/2e54ad6518a19738e54a640de4002e86c4ab.pdf.

[38] Han-Ping Chou, Jung-Ming Wang, Chiou-Shann Fuh, et al. "Automated lecture recording system". In: *2010 International Conference on System Science and Engineering*. Taipei: IEEE, 2010, pp. 167–172. ISBN: 9781424464746. DOI: 10.1109/ICSSE.2010.5551811.

[39] Christopher Brooks, Carrie Demmans Epp, Greg Logan, et al. "The who, what, when, and why of lecture capture". In: *ACM International Conference Proceeding Series* (2011), pp. 86–92. DOI: 10.1145/2090116.2090128.

[40] Hamad Odhabi and Lynn Nicks-Mccaleb. "Video recording lectures: Student and professor perspectives". In: *British Journal of Educational Technology* 42.2 (2011), pp. 327–336. ISSN: 00071013. DOI: 10.1111/j.1467-8535.2009.01011.x.

[41] Nenad Gligorić, Ana Uzelac, and Srdjan Krco. "Smart classroom: Real-time feedback on lecture quality". In: *2012 IEEE International Conference on Pervasive Computing and Communications Workshops, PERCOM Workshops 2012*. Lugano, Switzerland: IEEE, 2012, pp. 391–394. ISBN: 9781467309073. DOI: 10.1109/PerComW.2012.6197517.

[42] Michael Björn Winkler, Kai Michael Höver, Aristotelis Hadjakos, et al. "Automatic Camera Control for Tracking a Presenter during a Talk". In: *2012 IEEE International Symposium on Multimedia*. Irvine, CA, USA: IEEE, 2012, pp. 471–476. ISBN: 978-1-4673-4370-1. DOI: 10.1109/ISM.2012.96.

[43] Elisardo Gonzalez-Agulla, Jose L. Alba-Castro, Hector Canto, et al. "GaliTracker: Real-time lecturer-tracking for lecture capturing". In: *Proceedings - 2013 IEEE International Symposium on Multimedia, ISM 2013*. Anaheim: IEEE, 2013, pp. 462–467. ISBN: 9780769551401. DOI: 10.1109/ISM.2013.89.

[44] Nick Jones. "Quantification and substitution: The abstract space of virtual cinematography". In: *Animation* 8.3 (2013), pp. 253–266. ISSN: 17468477. DOI: 10.1177/1746847713508864. URL: http://anm.sagepub.com/cgi/doi/10.1177/1746847713508864.

[45] Takayuki Nagai, Toshiyuki Toyota, Takayuki Nagoya, et al. "Implementation of high-definition lecture recording system for daily use". In: *IEEE Global Engineering Education Conference, EDUCON*. Berlin: IEEE, 2013, pp. 520–525. ISBN: 9781467361101. DOI: 10.1109/EduCon.2013.6530155.

[46] Debosmit Ray. "Edge Detection in Digital Image Processing". In: *University of Washington: Department of Mathematics* (2013), pp. 1–11. URL: https://sites.math.washington.edu/~morrow/336_13/papers/debosmit.pdf.

[47] Andrews Sobral and Antoine Vacavant. "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos". In: *Computer Vision*

and Image Understanding 122 (2014), pp. 4–21. ISSN: 1090235X. DOI: 10.1016/j.cviu.2013.12.005. URL: http://ac.els-cdn.com.ezproxy.uct.ac.za/S1077314213002361/1-s2.0-S1077314213002361-main.pdf?_tid=e24f3caa-3e29-11e7-a6f2-00000aacb35e&acdnat=1495373707_432fbf0d4443d072f9493143d6f5ec6c.

[48] Benjamin Wulff, Alexander Fecke, Lisa Rupp, et al. "LectureSight: an open source system for automatic camera control for lecture recordings". In: *Interactive Technology and Smart Education* 11.3 (2014), pp. 184–200. ISSN: 17588510. DOI: 10.1108/ITSE-07-2014-0019. URL: https://doi.org/10.1108/ITSE-07-2014-0019.

[49] Yong Quan Chen, Chiung Fang Chang, and Po Chyi Su. "A tabletop lecture recording system based on gesture control". In: *2015 IEEE International Conference on Consumer Electronics - Taiwan, ICCE-TW 2015* (2015), pp. 372–373. DOI: 10.1109/ICCE-TW.2015.7216950.

[50] Dávid Cymbalák, Ondrej Kainz, and Jakab František. "Real-Time Automatic Selection of the Best Shot on Object in 4K Video Stream Based on Tracking Methods in Virtual Cropped Views". In: *International Journal of Computer and Electrical Engineering* 7.4 (2015), pp. 275–282. ISSN: 17938163. DOI: 10.17706/ijcee.2015.7.4.275-282. URL: http://www.ijcee.org/index.php?m=content&c=index&a=show&catid=76&id=1010.

[51] Hsien-Chou Liao, Ming-Ho Pan, Min-Chih Chang, et al. "An Automatic Lecture Recording System Using Pan-Tilt-Zoom Camera to Track Lecturer and Handwritten Data". In: *International Journal of Applied Science and Engineering* 13.1 (2015), pp. 1–18. ISSN: 1727-7841. DOI: 10.6703/IJASE.2015.13(1).1. URL: https://pdfs.semanticscholar.org/5660/9a1f2aa74960f5e65a71ca0b1bd56da38400.pdf.

[52] Johannes Ohlemacher. *Conception and Development of a Pipe & Filter Framework for C++*. Tech. rep. 2016, pp. 1–86.

[53] John Zhang and Tao Sun. "A Review of Some Local Feature Detection Algorithms". In: *John Zhang & Tao Sun International Journal of Image Processing* 103 (2016), pp. 2016–94. URL: http://www.cscjournals.org/manuscript/Journals/IJIP/Volume10/Issue3/IJIP-1071.pdf.

[54] Gagan Baraskar and Pooja Thakre. "Evaluation of Canny and Sobel Edge Detection Technique using Xilinx System Generator". In: *International Journal of Scientific Research in Science and Technology* 3.2 (2017), pp. 53–56. ISSN: 2395-6011. URL: http://ijsrst.com/paper/872.pdf.

[55] Oussama H. Hamid. "Spike-Like Temporal Patterns Exhibit Persistent Frequency and Intensity of Viewing Behaviour towards Video-Recorded Lectures Published on YouTube ". In: *Internet Technologies and Applications*. Wrexham, UK: IEEE, Nov. 2017, pp. 173–176. ISBN: 978-1-5090-4815-1. URL: https://ieeexplore-ieee-org.ezproxy.uct.ac.za/stamp/stamp.jsp?tp=&arnumber=8101932.

[56] Md Nazrul Islam, Manjeevan Seera, and Chu Kiong Loo. "A robust incremental clustering-based facial feature tracking". In: *Applied Soft Computing Journal* 53 (2017), pp. 34–44. ISSN: 15684946. DOI: 10.1016/j.asoc.2016.12.033. URL: http://dx.doi.org/10.1016/j.asoc.2016.12.033.

[57] Moneish Kumar, Vineet Gandhi, Remi Ronfard, et al. "Zooming On All Actors: Automatic Focus+Context Split Screen Video Generation". In: *Computer Graphics Forum* 36.2 (2017), pp. 455–465. ISSN: 14678659. DOI: 10.1111/cgf.13140. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13140.

[58] Yunseong Lee, Alberto Scolari Politecnico, Di Milano, et al. *Open access to the Proceedings of the 13th USENIX Symposium on Operating Systems Design and Implementation is sponsored by USENIX Pretzel: Opening the Black Box of Machine Learning Prediction Serving Systems PRETZEL: Opening the Black Box of Machine Learn*. 2018, pp. 611–626. ISBN: 978-1-939133-08-3. URL: https://www.usenix.org/conference/osdi18/presentation/lee.

[59] Tong Liu, Shakeel Alibhai, Jinzhen Wang, et al. "Exploring Transfer Learning to Reduce Training Overhead of HPC Data in Machine Learning". In: *2019 IEEE International Conference on Networking, Architecture and Storage (NAS)*. EnShi, China: IEEE, 2019, pp. 1–7. DOI: 10.1109/NAS.2019.8834723. URL: https://ieeexplore-ieee-org.ezproxy.uct.ac.za/stamp/stamp.jsp?tp=&arnumber=8834723.

[60] Cynthia Rudin. *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*. Tech. rep. Duke University, 2019, pp. 206–215.

[61] Marco Furini, Giovanna Galli, and Maria Cristiana Martini. "An Online Education System to Produce and Distribute Video Lectures". In: *Mobile Networks and Applications* 25.3 (June 2020), pp. 969–976. ISSN: 15728153. DOI: 10.1007/S11036-019-01236-4/TABLES/2. URL: https://link-springer-com.ezproxy.uct.ac.za/article/10.1007/s11036-019-01236-4.