Low Resource, Post-processed Lecture Recording from 4K Video Streams

Charles Fitzhenry[1], Tanweer Khatieb[1], Patrick Marais[1], Stephen Marquard[12]

[1]Dept. of Computer Science, University of Cape Town, Cape Town, South Africa

[2]Centre for Innovation in Learning and Teaching (CILT).

Centre for Innovation in Learning and Teaching

SAICSIT 2022, Cape Town

19 July 2022

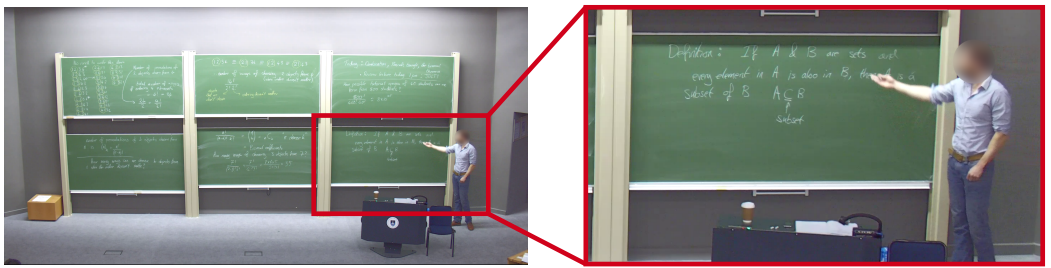
Introduction

- High-resolution 4K cameras are increasingly used for lecture recording as they are affordable and allow for easy reading of board/screen context and capture the full stage area.
- Filesize over 2GB for a 45-60 minute lecture, mitigating the use of such technology in a low-resource environment.
- Developed a low resource 4K lecture recording solution, which addresses these problems through a computationally efficient video processing pipeline with a fixed memory footprint that won't be exceeded.
- Extension to open source system *TRACK4K* [1]and will also be made available under the same GitHub profile once completed.

[1] https://github.com/LectureTracking/trackhd

Introduction

- Front-end uses very low cost algorithms to segment presenter motion and writing/board surfaces from the stream.
- Back-end serves as a virtual cinematographer (VC), combining this contextual information to draw attention to the lecturer and relevant content.
- Significantly reduces file size by over 80% while maintaining writing clarity.



Related Work

- *LectureSight*[2] open source live camera tracking system for lecture recording that requires expensive Pan-Tilt-Zoom (PTZ) cameras.
- Similar to our work, Cymbalák et al.[3] extracts a smaller crop region from multiple 4K camera streams to select the best shot. Real-time tracking approach and not specialised for lecture recording and does not consider board usage.

[2] Benjamin Wulff and Alexander Fecke. "LectureSight - An open source system for automatic camera control in lecture recordings". In: *2012 IEEE International Symposium on Multimedia*. Irvine, CA, USA: IEEE, 2012, pp. 461–466.

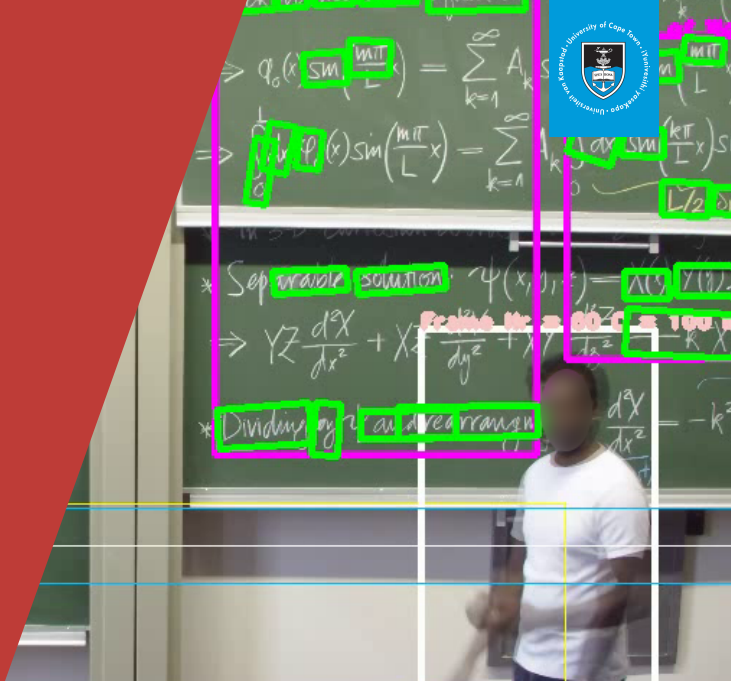
[3] Dávid Cymbalák et al. "Real-Time Automatic Selection of the Best Shot on Object in 4K Video Stream Based on Tracking Methods in Virtual Cropped Views". In: *International Journal of Computer and Electrical Engineering* 7.4 (2015), pp. 275–282.

Diversity of Venue Configurations

- Wide range of camera angles and venue/board configurations.

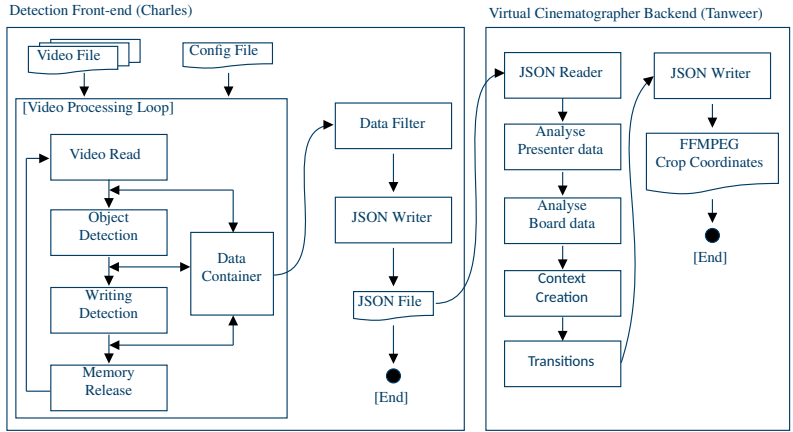
Presenter & Writing Detection Front-end

Charles Fitzhenry



System Overview

- Overview of the complete system architecture.



Presenter Detection Overview

- Machine learning based approaches very successful, however, typically inefficient without GPU acceleration.
- E.g *OpenPose* [4] inference time GPU: 36ms vs. CPU: 10396ms.
- Presenter detection requires a higher detection frequency to cope with a moving presenter. This means a DL approach infeasible for our context, if used without a GPU.

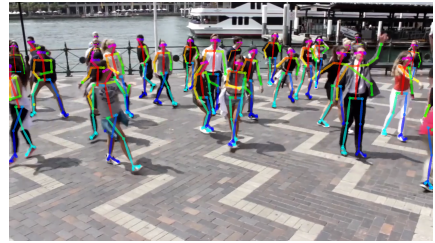


Image source: *OpenPose* [4] paper

[4] Zhe Cao et al. "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.1 (July 2019), pp. 172–186.

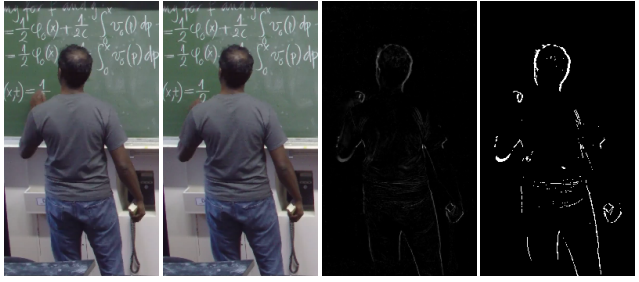
Presenter Detection Overview

- Having surveyed the literature, temporal frame differencing was identified as a very low cost motion detection method and we explore the feasibility of using it for presenter detection in our context of driving a VC.
- Many object detection applications require a tight bounding box, however, for our purpose we do not require this precision — we also developed heuristics to mitigate weaknesses of frame differencing.
- Carefully designed a processing pipeline capable of processing videos in segments based on RAM availability — enabling the use of low end machines for processing without requiring expensive GPUs.
- Used *C++* and image processing algorithms from the *OpenCV* [5] library.

[5] www.opencv.org

Motion Detection

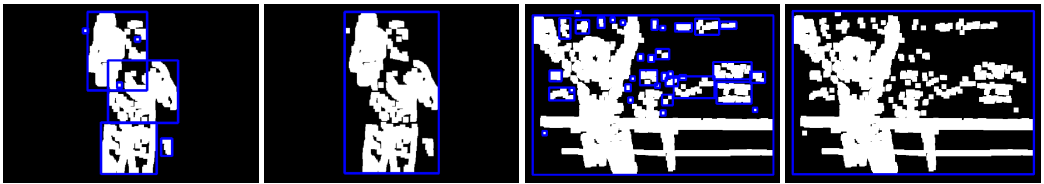
- At the core, we use pixel-wise frame differencing to detect motion.
- Additional frames are skipped to further reduce computational cost.
- Further processed with adaptive thresholding, morphology and connected components to be robust under various lighting conditions.
- Detected objects are enclosed in bounding rectangles.



(a) Frame A (b) Frame B (c) Difference (d) Threshold

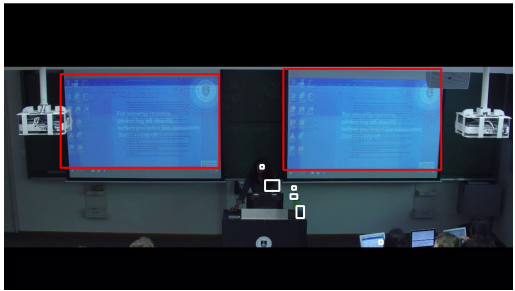
Merging Bounding Boxes

- Merge overlapping bounding boxes.
- Board motion or students walking past will cause oversized bounding boxes — although this is transient and not a concern.

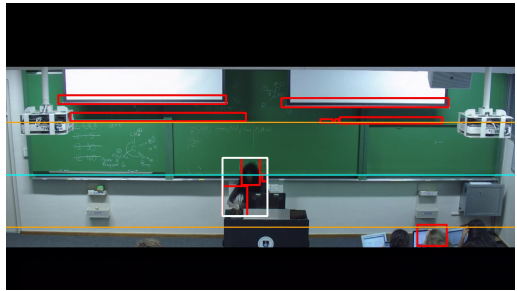


Filter Heuristics

- Set of heuristics developed to mitigate the limitations of temporal differencing.
- Object area ratio ($\frac{\text{object area}}{\text{frame area}}$) and aspect ratio ($\frac{x}{y}$) are used to detected and remove anomalies.



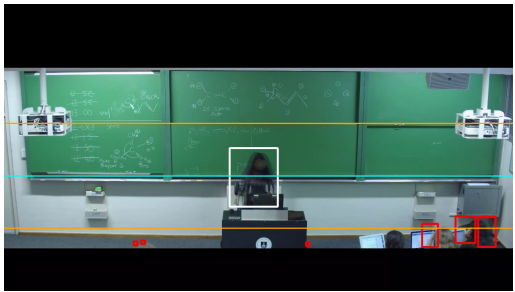
(a) Area Ratio



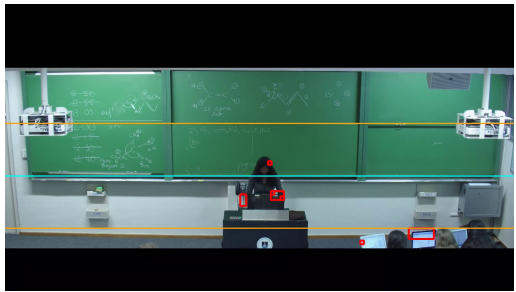
(b) Aspect Ratio

Filter Heuristics

- Vertical limits created using the mean of all bounding box y-coordinates.
- Remove any remaining boxes not within ± 1 SD (yellow lines).
- Stationary presenter produces no valid bounding box — So we interpolate last known box to future reappearance only if they intersect.



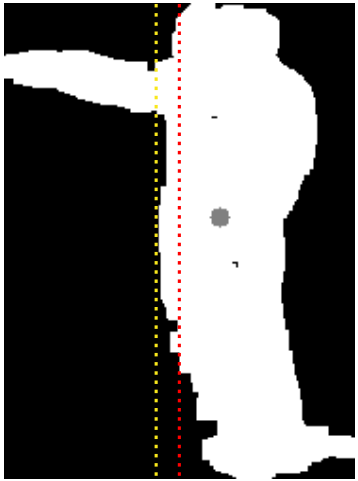
(a) Vertical Limits



(b) Interpolation

Gesture Detection Feasibility

- Explored detecting left/right gestures as an extension to presenter detection to provide additional context in conjunction with board usage.
- Analysed the spatial distribution of pixels in the bounding box — any unbalance beyond a set threshold would indicate a gesture.
- Unfortunately, the quality of silhouettes produced by the differencing approach were not sufficient for this purpose.



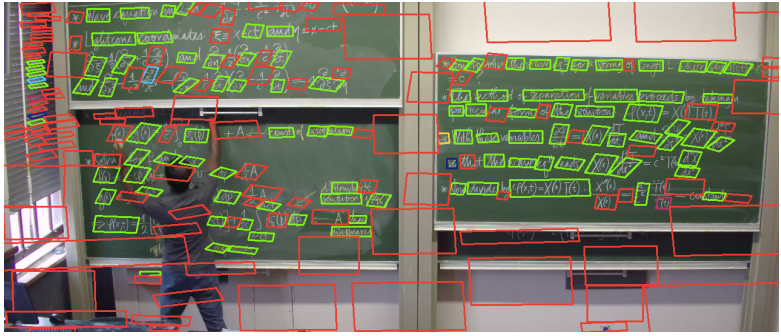
Writing Detection

- Writing detection aims to assist the VC in making context-centred framing decisions by including boards/screens relevant to the presentation.
- Writing is added incrementally over time and remains in the same location for longer.
- Using a lower detection frequency (once every 30 seconds) to avoid adding much computational overhead, enables us to explore a machine learning based approach without using a GPU.
- At the time of our literature survey, the Efficient and Accurate Scene Text (EAST) detector[6] was considered robust and state-of-the-art.
- Instead of developing a specialised detector (which is a separate field of research), we explore the feasibility of using EAST with their pre-trained neural network purely as a proxy for finding board/screen usage regions.

[6] Xinyu Zhou et al. "EAST: An Efficient and Accurate Scene Text Detector". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, USA: IEEE, July 2017, pp. 2642–2651.

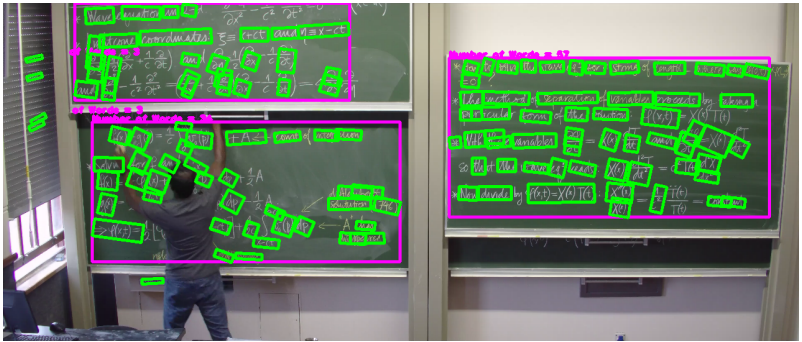
EAST Text Detector

- Output produced by EAST and colour coded confidence scores for the detection.
- Detection (words) are enclosed in rotated rectangles.



Clustering Words

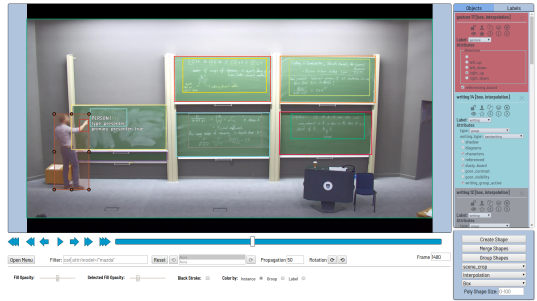
- Words are clustered to create regions of writing that may be related to the presentation.
- Clustering used the vertices of the rotated rectangle enclosing each word.



Evaluation and Results for the *Front-end*

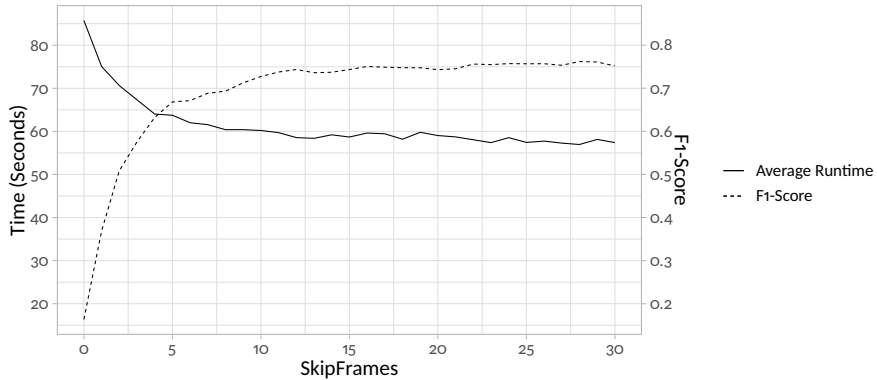
Ground Truth Data

- Annotated frames to use as GT for evaluating the front-end accuracy.
- Evaluated detected boxes against annotated ground truth bounding boxes.
- Calculated Intersection over Union (IOU) between our detections (yellow) and GT (green).
- Two examples of cases that will be considered True Positive with an IOU threshold of 0.3.



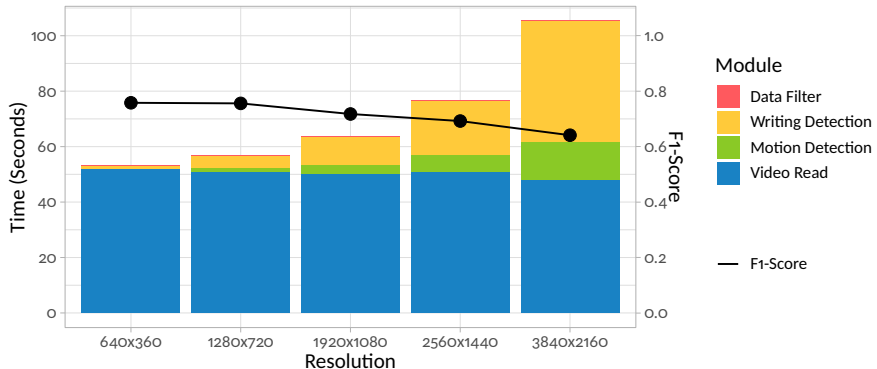
Presenter Detection Runtime and Accuracy

- Effects of *Skip Frame* parameter vs. overall front-end runtime.
- Trended downward with $SF > 28$



Front-end Module Runtime and Accuracy

- Resolution vs. writing detection runtime and F1-Score.
- At 720p an unavoidable $\approx 90\%$ used for video read & decoding and a small amount for frame resize operation — Solid-State Drive (SSD) could improve these times.



Writing Detection and Accuracy

F1-Scores were calculated using the method of Wolf and Jolion[7] to accommodate fragmented detections.

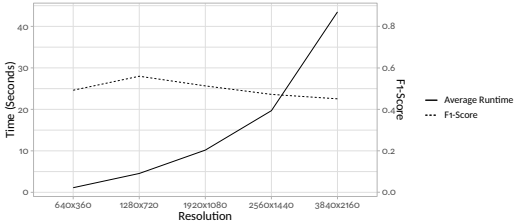


Figure: Resolution vs. writing detection runtime and F1-Score

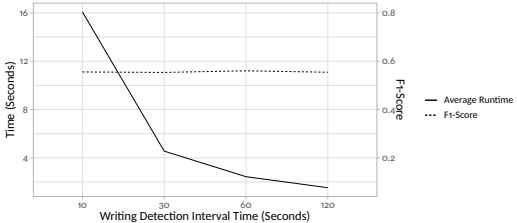


Figure: Writing detection interval vs. writing detection runtime and F1-Score.

[7] Christian Wolf and Jean Michel Jolion. "Object count/area graphs for the evaluation of object detection and segmentation algorithms". In: *International Journal of Document Analysis and Recognition (IJDAR)* 8.4 (Sept. 2006), pp. 280–296.

Environmental Effects on Presenter Detection

- Student heads in view had almost no effect on our detection with filter on.
- Scenes with multiple objects performed the worst — although not a concern since events are typically transient.

Attribute	Value	F1 (Filter off)	F1 (Filter on)
Overall	NA	0.55	0.78
Lighting	Low	0.39	0.70
	Good	0.64	0.82
Student Heads	No	0.60	0.79
	Yes	0.46	0.74
Multiple Objects	No	0.56	0.80
	Yes	0.38	0.43
Presenter Motion	Low	0.36	0.68
	High	0.74	0.88

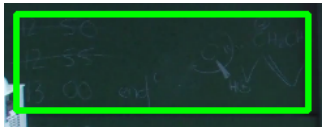
Environmental Effects on Writing Detection

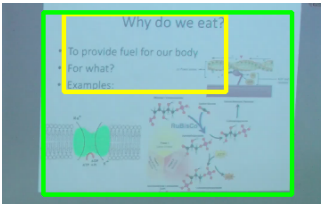
- Overall *F1-Score* $= 0.55$
- Poor contrast includes dusty boards, low lighting and poor chalk/text contrast.
- Diagrams are a known limitation of using EAST
- Text is detected as a means general board usage detection, so a high accuracy not essential.

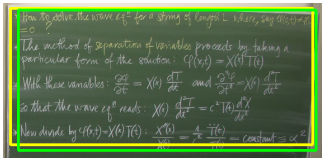
Attribute	Value	F1
Contrast	Good	0.62
	Poor	0.19
Diagrams	Absent	0.62
	Present	0.39
Text	Handwriting	0.51
	Projector	0.57

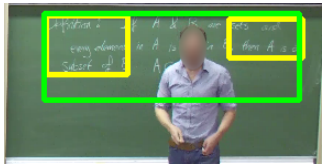
Environmental Effects on Writing Detection

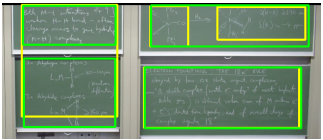
- Observed that many detected boxes (yellow) are smaller than the GT (green) boxes. The GT data only provides a single tightest bounding box enclosing all content on each board.









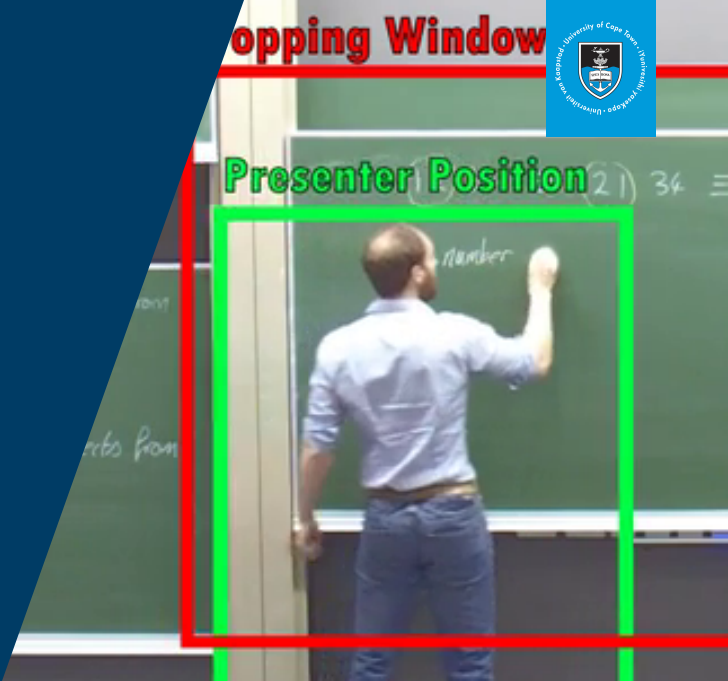


Summary

- Our low-cost temporal differencing in conjunction with our filter heuristics have a very small processing time (excluding disk IO) and is a feasible alternative, to more complex solutions, for the purpose of presenter detection.
- Using the EAST detector at a low frequency is a feasible approach in terms of processing speed, however, its accuracy is limited by frames with images or diagrams.

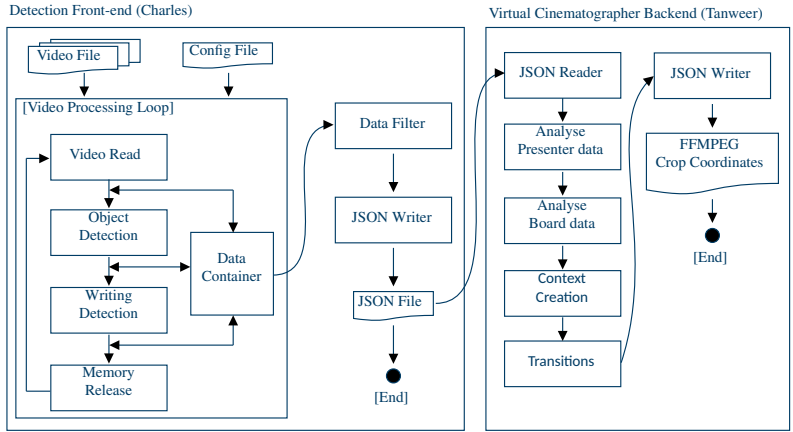
Virtual Cinematographer Back-end

Tanweer Khatieb



System Overview

- Overview of the complete system architecture.



What is the Virtual Cinematographer?

The Virtual Cinematographer (VC) has the following functions in the system:

1. Takes input from the front-end (JSON format)
2. Builds internal representation
3. Performs scene analysis
4. Makes framing decisions
5. Produces output instructions (JSON format) for FFMPEG

What is the Virtual Cinematographer?

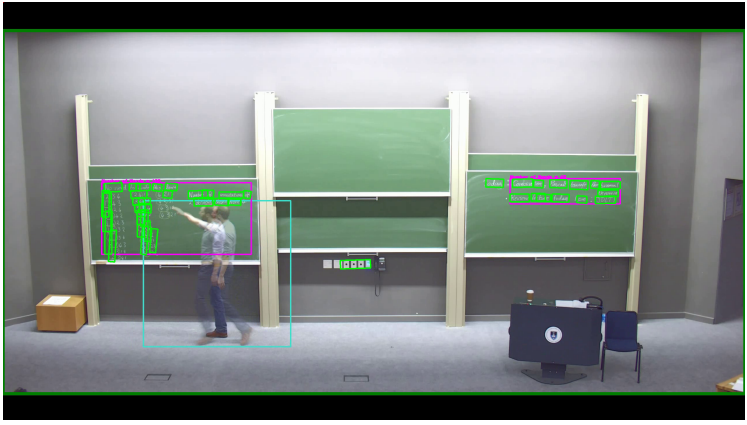
The VC follows some heuristics/guidelines when making framing decisions such as:

- Guide the viewer's attention to what is important
- Move between shots smoothly
- Only move the camera when necessary
- Do not make jump cuts
- etc.

VC Input

Labeled data from the front-end helps the VC to make decisions like:

- How best to frame the content
- When to make a transition



VC Input

The VC expects the following data from the front-end:

- Presenter‘(s‘) bounding box(es) for every frame
- Board bounding boxes
- Board features
- Board feature count
- Number of frames in video
- etc.

VC Internal Representation

The VC creates an internal model of the input data in memory, which has the following benefits:

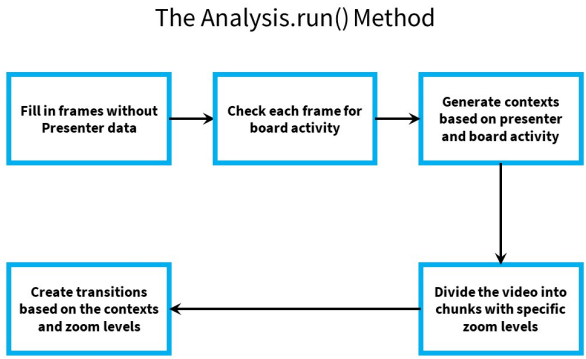
- Reduces run time by keeping away from File I/O
  - Having data in memory allows quicker reading and writing
  - Allows multiple passes over the data without the time overhead
- Non-destructive manipulation of input data
  - Keeping data in memory allows the VC to make copies, overwrite, adjust, or otherwise modify the data without affecting the input file

VC Scene Analysis

The most important step of the VC Module's runtime is to perform scene analysis

- Provides context for events
  - Allows events to be separated in terms of these contexts
- Good understanding of events leads to better decision-making
  - Better decision-making leads to better framing and video quality

VC Scene Analysis - Overview

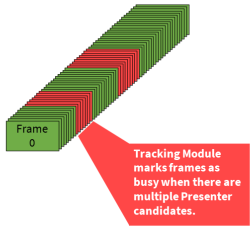


The Analysis.run() method performs the scene analysis for the VC

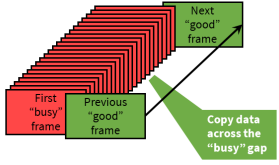
VC Scene Analysis - Fixing ‘gaps’ in the data

A common issue for the VC is an incomplete sequence of presenter bounding boxes.

- On frames with a lot of movement presenter data is not saved.
- These ‘gaps’ need to be filled
- VC duplicates the last known position of the presenter into all frames in the ‘gap’



The VC corrects “busy” frames by saving the last known position of the presenter before the patch of “busy” frames into the first frame after the patch.



VC Scene Analysis - Boards and Board Usage

Board information helps identify what is important in the venue

- Content on boards are linked to what is being explained
- Boards being modified indicate activity
- Bounding box of the used board helps determine the limits of the context
  - VC knows to include presenter and board in the frame
  - Sets limits of the Cropping window to include both bounding boxes

VC Scene Analysis - Board Usage Detection

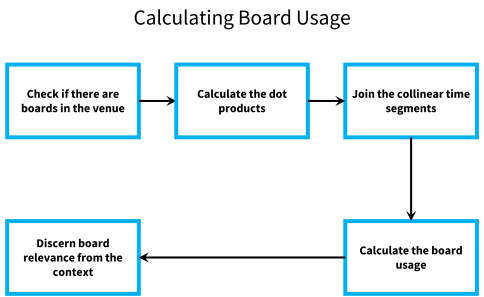
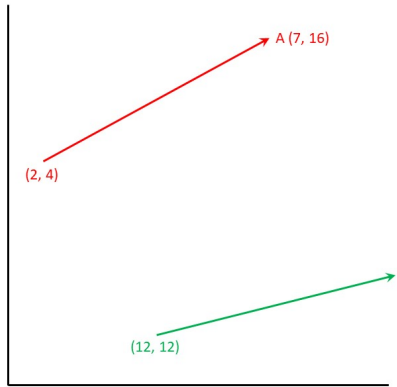


Figure: Overview of the board usage detection algorithm

Board Usage Detection - What do you mean by Dot Product?



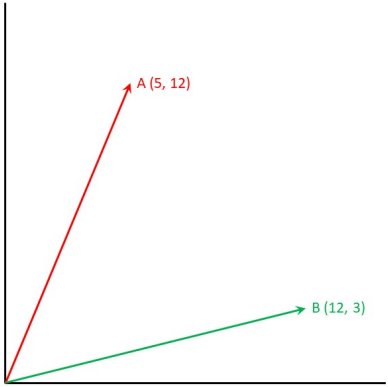
We can get the angle between these two vectors using the dot product.

The first step is to represent the vectors mathematically, which is done by bringing their starting points to the origin.

For Vector A, we subtract 2 on the x-axis and 4 from the y-axis. For Vector B, we subtract 12 from both the x and y axes.

The Dot Product Explained

Board Usage Detection - What do you mean by Dot Product?



Once we have moved the vectors to the origin, we can represent the vectors using only the end point.

Vector A now becomes A(5, 12), and Vector B becomes B(12, 3).

The Dot product can now be written as:

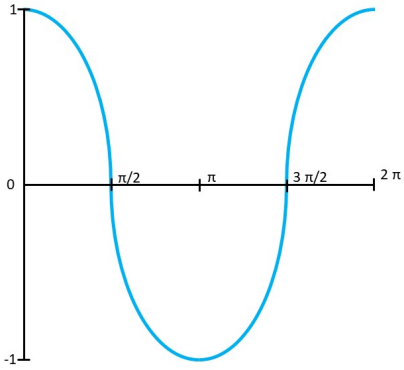
$$A.B = |A||B| \cos \theta$$

Now, since we want to know Cos θ, we can rewrite this formula to solve for it.

The Dot Product Explained

Board Usage Detection - What do you mean by Dot Product?

The Dot Product Explained



$$A.B = |A||B| \, Cos(\theta)$$

$$\therefore \text{Cos}\,(\theta) = \frac{A.B}{|A||B|}$$

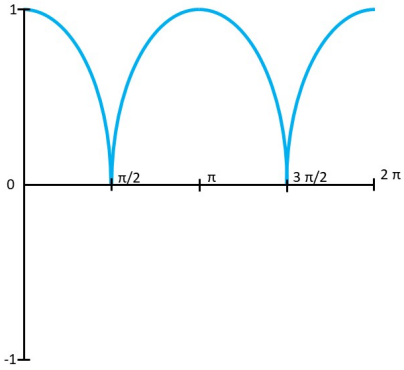
Cos(θ) = 1 means parallel lines in the same direction

Cos(θ) = -1 means parallel lines in opposite directions

Cos(θ) = 0 means perpendicular lines

Board Usage Detection - What do you mean by Dot Product?

The Dot Product Explained

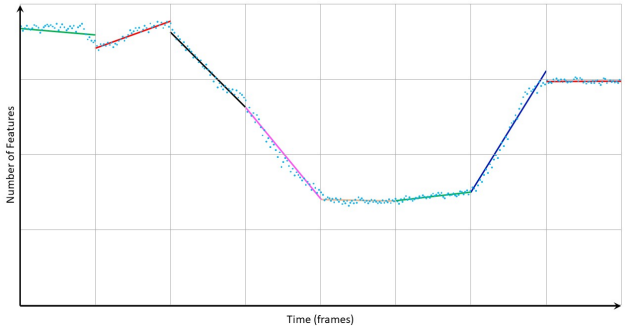


We can use the absolute value if we are not concerned about direction.

$$\therefore |\text{Cos}\ (\theta)| = \left|\frac{A.B}{|A||B|}\right|$$

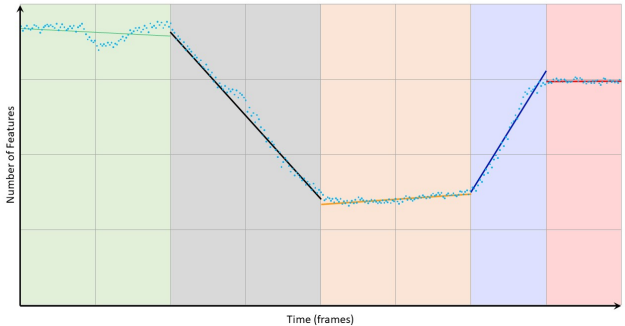
|Cos(θ)| = 1 means parallel lines in the same direction, and |Cos(θ)| = 0 means perpendicular lines.

Board Usage Detection - Dot Product in use



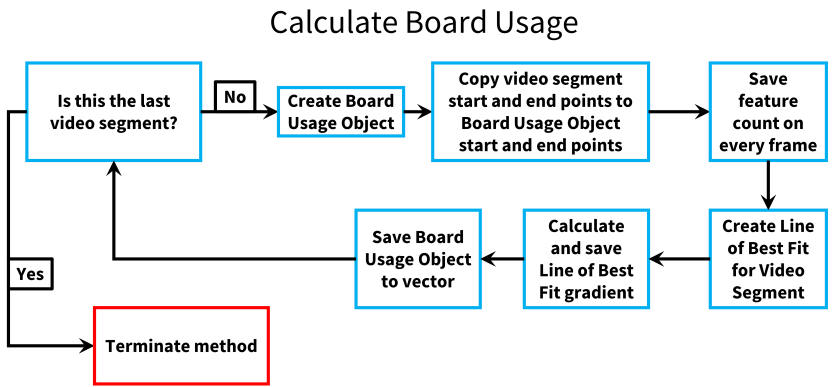
Example: video broken into equal time segments

Board Usage Detection - Dot Product in use



Example: Colinear segments grouped and lines are joined

VC Scene Analysis - Board Usage Detection



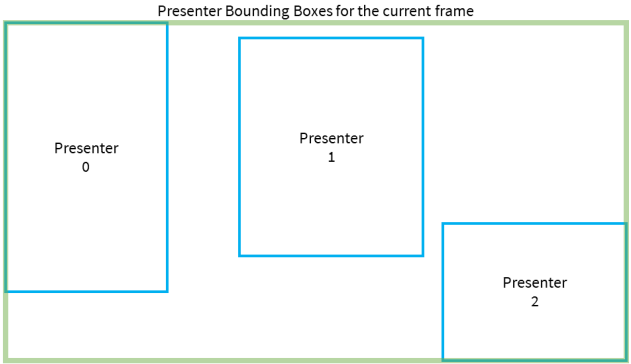
VC Scene Analysis - Contexts

After the VC has the following data:

- All presenter bounding boxes in all frames
- All board bounding boxes in all frames
- All board usages in all frames

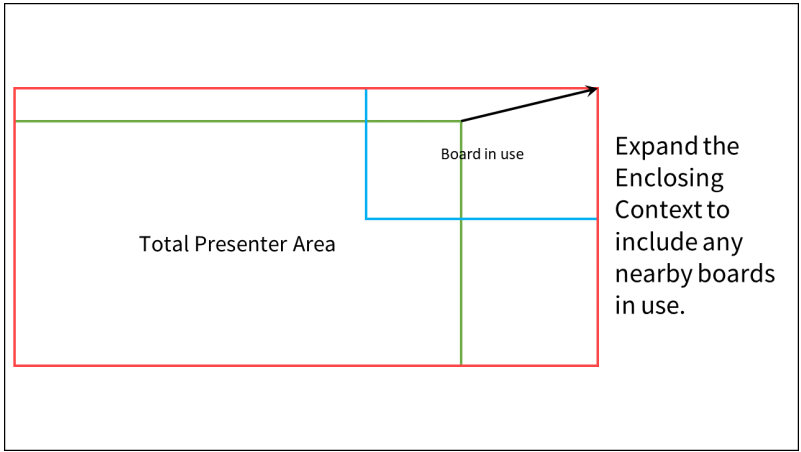
The VC then groups this information into bounding boxes called ‘Enclosing Contexts’ (using the above data) which are then used to determine framing choices and transitions

VC Framing Decisions - Enclosing Contexts



The VC bounds all the presenter's bounding boxes into a single bounding box based on the extremes in all 4 directions.

VC Framing Decisions - Enclosing Contexts



VC Framing Decisions - Connecting Enclosing Contexts

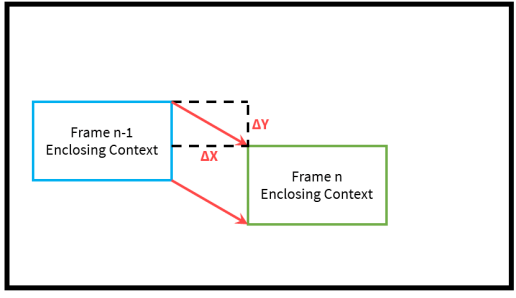
Once the Enclosing Context on each frame is identified the VC must decide:

- How to deal with each Enclosing Context separately
- How to group neighbouring Enclosing Contexts into chunks
- How to join neighbouring chunks with transitions

To manage this, the VC does the following:

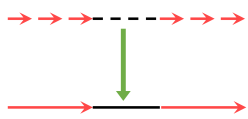
- Check Enclosing Context motion
- Combine consistent motion & correct aspect ratios
- Group Enclosing Contexts into chunks
- Generate Transitions

VC Framing Decisions - Check Enclosing Context Motion



The VC must evaluate the motion of the Enclosing Contexts across neighbouring frames by comparing their positions.

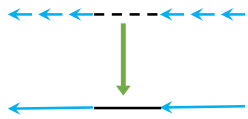
VC Framing Decisions - Combine Consistent Motion



Smaller movements to the right are combined

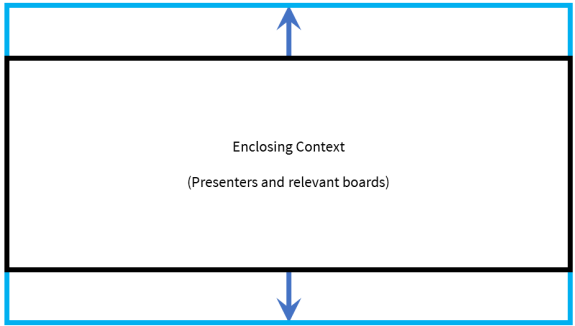
Diagonal and horizontal movements to the right are collapsed

Smaller movements to the left are combined



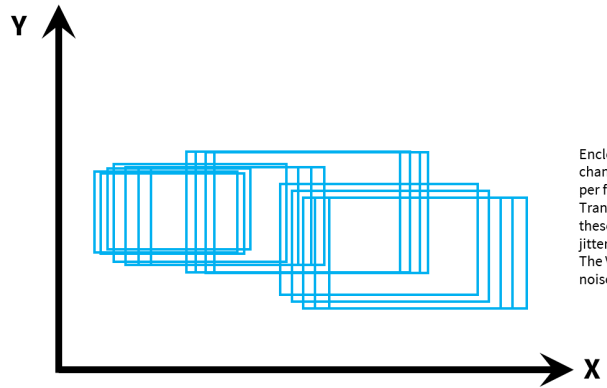
Diagonal and horizontal movements to the left are collapsed

VC Framing Decisions - Correct Aspect Ratios



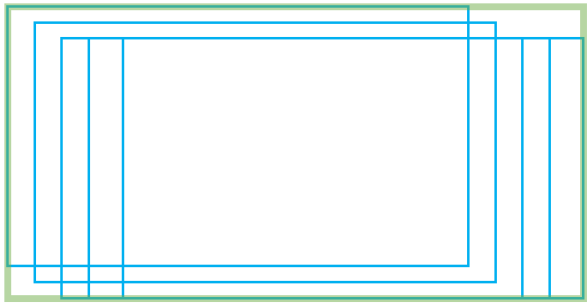
The Enclosing Context must be expanded (horizontally or vertically) to match the aspect ratio of the output video.

VC Framing Decisions - Group Enclosing Contexts into Chunks



Enclosing Contexts change size and position per frame.
Transitions between these would be too jittery.
The VC must reduce the noise.

VC Framing Decisions - Group Enclosing Contexts into Chunks

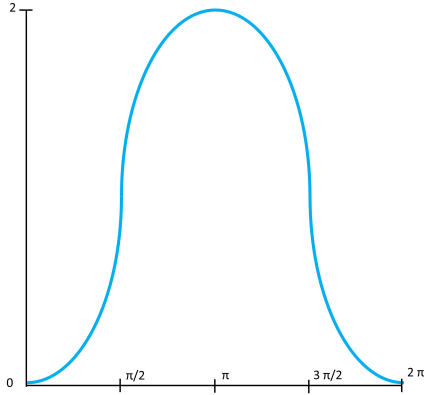


The VC groups the Enclosing Contexts of neighbouring frames based on their size and proximity.

VC Framing Decisions - Generate Transitions between chunks

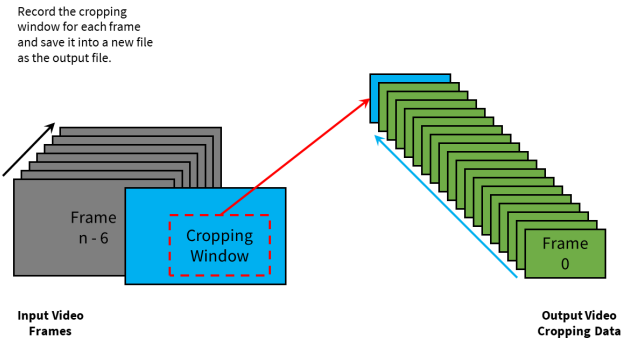
The VC must transition smoothly from one chunk to the next.

- The VC uses the function $\cos(x + \pi) + 1$ with domain $[0, 2\pi]$ radians.
- Applied on four movements:
  1. Top Left corner Horizontal motion
  2. Top Left corner Vertical motion
  3. Bottom Right corner Horizontal motion
  4. Bottom Right corner Vertical motion
- These 4 movements allow pan & zoom simultaneously



VC Output

The Cropping Window is saved as the final Enclosing Context for each frame and is then written to JSON



Evaluation and Results for the *Virtual Cinematographer*

Evaluating the Virtual Cinematographer

To Test the VC we needed to perform a user evaluation on the following aspects:

- How did different configurations of the VC affect user reception?
- How did different lecture venue layouts affect the VC's performance?

VC Evaluation Layout

We set up a collection of video pairs with videos placed side by side and participants would choose their preferred video from each pair

- Each video clip in a pair had the same:
  - Input video
  - Time interval
  - Venue Layout
- Each video could play independently of the other and was set to play on loop
- The whole evaluation was set up to include no more than 1 hour of collective video clip time
  - Prevents participants become bored and tired
  - Minimises inconvenience to participants

VC Evaluation Layout

Please watch the video clips and choose which of them you preferred in the options below:

- ◯ A. Left Video
- ◯ B. Right Video

Reset Selection

VC Evaluation Layout

After choosing their preferred video in a pair, they would have to answer the 8 questions in this matrix

- All questions phrased as affirmations
- Gave us quantity to the quality-based evaluation
- Each video pair was followed by this matrix

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

VC Evaluation Layout - Configuration Types

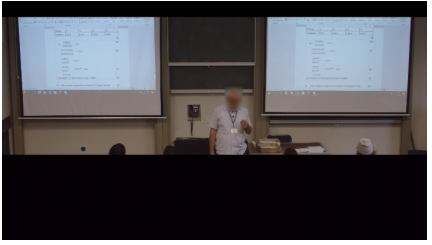
There were 3 different configuration types used in the evaluation:

- Configuration 1 - ‘Low Laziness’
- Configuration 2 - ‘High Laziness’
- Control - No VC intervention at all

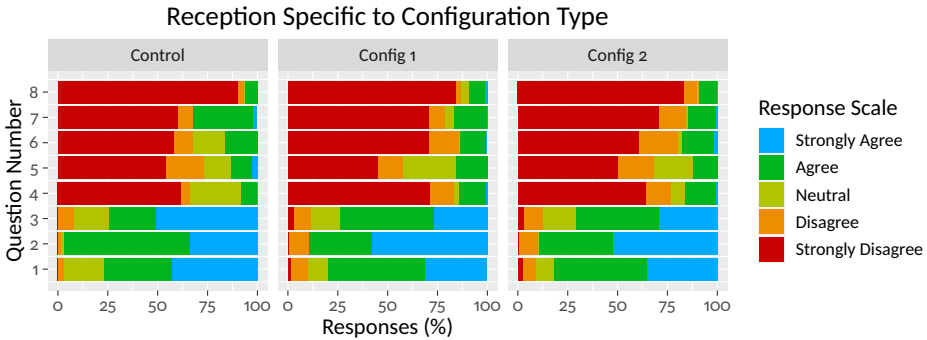
The ‘laziness’ here refers to the sensitivity with which the VC applies transitions

- High Laziness - less responsive and fewer transitions
- Low Laziness - more responsive and more transitions

VC Evaluation Layout - Venue Layouts

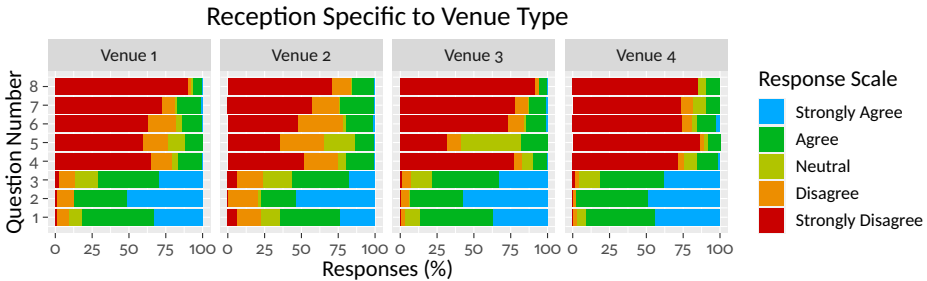


VC Evaluation Results - Configuration Type



Throughout the evaluation we observed a positive reception for the first 3 matrix questions and a negative reception for the remaining questions

VC Evaluation Results - Venue Layout



As with the configuration types above, we observed a positive reception for the first 3 matrix questions and a negative reception for the remaining questions

VC Results - Final Observations

From the evaluation, regardless of configurations or venue layouts, we can observe the following about the VC:

- Transitions were smooth
- Participants could see what they wanted to watch
- Participants were satisfied with transition frequency

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

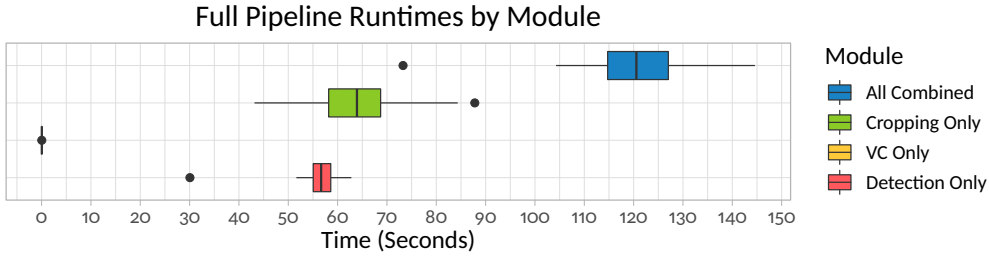
VC Results - Final Observations

What else can we observe from results?

- The questions with negative responses were specific to varying factors in a venue
- The VC is fit for purpose since it fulfills the 3 main criteria (tested in the first 3 questions of the matrix)
    1. Smooth transitions
    2. Shows relevant information
    3. Transitions only when appropriate
- We also observe that most participants had little experience with watching lecture videos, which may contribute to the negative reception in the last 5 questions

Full Pipeline Runtimes

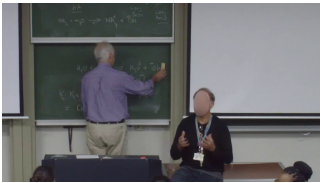
- Distribution of runtime per module on a sample set ($n = 88$) of 2-minute video clips.
- A mean filesize reduction of 81.3% from the 4K video to the cropped 720p video.
- VC only takes an average of 40.3 ms to run a 2-minute test video — low disk IO.
- < 145 seconds to fully process a 2-minute video (or 1.5x input video duration)
- *ffmpeg* [8] filters used for cropping and rescaling the video frames.



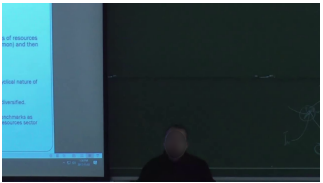
[8] https://ffmpeg.org/ffmpeg-filters.html

Full Pipeline — Video Examples

(a) 00:44



(b) 01:17



(c) 00:44

Limitations and Future Work

- Could explore a ML based presenter and gesture detection at a lower frequency similar to writing detection — can’t be too infrequent else gestures will be skipped.
- Support ability to detect figures/diagrams to detect full board usage.
- Further VC user studies with more experienced lecture audiences could be conducted to obtain more conclusive results.

Acknowledgements

We would like to thank the following people and institutions for their contributions to our project:

- Department of Computer Science & Centre for Innovation in Learning and Teaching (CILT) at the University of Cape Town.
- Lecturers appearing in the videos.
- Participants in VC evaluation.
- Data Annotators for ground truths in testing the front-end.
- Our families.

Thank You!
Questions?