Smart Multimodal Classroom Video Recorder

Bhavya Surana, Grace Chong, William Clavier

March 2, 2024

Abstract

This project proposes a smart multimodal classroom video recording system that automatically composes multiple content streams—camera feeds, slides, and whiteboard—based on real-time cues like gestures and spoken references. By leveraging computer vision, automatic speech recognition (ASR), and content analysis, it can dynamically pan, zoom, and switch between sources to create a more engaging, context-aware lecture recording. The goal is to overcome the limitations of static cameras and provide a richer, more immersive experience for both live and recorded viewers.

Introduction

As more educational institutions adopt lecture capture systems, the conventional setup often relies on a single static camera positioned at the back of the room, capturing both the instructor and the presentation screen in a wide-angle shot. While this method offers a basic overview and fulfills the basic requirement of recording the lecture, it fails to capture the nuanced gestures, facial expressions, or detailed whiteboard interactions that frequently comprise a significant portion of classroom discourse. Consequently, the static, one-size-fits-all video model does not convey the richness of the live classroom environment.

Instructors frequently rely on multiple sources—such as PowerPoint slides, external videos, digital annotations, whiteboard writing, and physical demonstrations—making it difficult for students viewing the recording to fully grasp context shifts and instructor cues. For example, if an instructor verbally references a slide or points to a specific diagram on the board, a single wide shot may not highlight these crucial moments effectively. This can leave viewers with an obscured perspective and cause important references to be lost in translation.

In this project, we aim to address these limitations by building a "virtual video director" that composes various input streams in real-time. By leveraging computer vision, automatic speech recognition (ASR), and multimodal content analysis, the system will detect relevant visual and auditory cues—such as when the instructor points out slide content, draws on the whiteboard, or shifts focus between lecturing and student engagement—and autonomously respond with appropriate framing. By integrating multiple content sources (e.g., a wide-angle or pan-tilt-zoom camera, screen capture, and audio), the system can selectively zoom, pan, and overlay to produce a dynamic, context-aware recording. This project proposal outlines the design, implementation, and evaluation of this system, using existing lecture recordings (such as DS542 classroom Echo360 captures) to demonstrate the potential of an adaptive lecture capture approach.

Related Work

Automated Lecture Capture Systems

Commercial platforms like Panopto and Echo360 (used at Boston University) provide hardware and software solutions for lecture capture by recording instructors and their presentation slides—usually from a single camera angle [1], [2]. Although these solutions improve accessibility, they often rely heavily on either manual switching or rudimentary motion-tracking, resulting in static or inflexible recordings when instructors or presentation media change positions.

Intelligent Camera Control and Gesture Recognition Research on intelligent camera control for classroom environments has been attempted. Emerging frameworks like OpenPose, MediaPipe, and Datavideo detect motion or teacher location to guide pan-tilt-zoom cameras automatically [3]. However, these still lack robust gesture recognition or deep content understanding—elements that could trigger more context-aware transitions.

Speech and Content Analysis

Parallel advances in automatic speech recognition (ASR) have improved real-time transcription and keyword detection, supporting enhanced lecture recordings [4]. Beyond transcription, research on content analysis—including optical character recognition (OCR) on slides—can detect slide transitions or identify specific diagram references [5]. Integrating these tools can automatically match an instructor's spoken words with precise slide elements, making it possible to adapt camera views at the exact moment relevant content is mentioned.

Multimodal Fusion Approaches

Recent studies emphasize multimodal fusion, combining video, audio, and textual cues to better understand classroom activities [6]. A framework incorporating computer vision, ASR, and content analysis can deliver more engaging and context-aware lecture recordings.

Proposed Work

Our proposed solution is a real-time intelligent video composition system that integrates deep learning techniques to analyze multimodal classroom content and automatically produce engaging, context-aware recordings. The system architecture consists of the following components:

Video Analysis Pipeline

- Person Detection and Tracking: We will implement a YOLOv8-based model to detect and track the instructor's movements across the classroom, allowing our system to follow them as they move between presentation areas.
- Pose Estimation: Using MediaPipe or a similar framework, we'll detect the instructor's gestures and body language to infer teaching intent (e.g., pointing at slides, writing on the board).
- Activity Recognition: We'll train a custom action recognition model using a 3D CNN architecture (like I3D or SlowFast) to classify instructor activities (lecturing, writing, demonstrating).

Content Understanding

- Presentation Content Analysis: We'll implement OCR (using Google's Vision API or Tesseract) to extract text from presentation slides and develop a content classifier to categorize slide types (title, bullet points, diagrams, code).
- Whiteboard/Chalkboard Detection: A dedicated vision model will detect and segment whiteboard/chalkboard regions, with change detection to identify when new content is added.
- Audio Speech Recognition: Using Whisper or a similar ASR system, we'll transcribe instructor speech in real-time to identify content references and teaching cues.

Multimodal Integration and Decision Engine

- Context Fusion Module: We'll develop a transformer-based architecture to integrate signals from video, slides, and audio, producing a unified understanding of the classroom context.
- Shot Selection Policy: Using reinforcement learning, we'll train a policy network that decides optimal camera framing and content composition based on the fused context.
- Smoothing and Transition Logic: To ensure professional-quality video output, we'll

implement temporal smoothing and intelligent transitions between different composition states.

Implementation and Integration

- OBS Integration: We'll develop a plugin for Open Broadcaster Software that implements our intelligent composition system, allowing real-time control of scenes, transitions, and overlays.
- Resource Optimization: We'll employ model quantization and hardware acceleration to ensure the system can operate in real-time on standard classroom computing hardware.

Our approach is promising because it combines state-of-the-art techniques from computer vision, natural language processing, and multimodal learning into a cohesive system specifically designed for the educational context. Unlike generic video production systems, our solution will understand the pedagogical flow of a lecture and make composition decisions that enhance the learning experience. The modular architecture also allows for progressive improvements to individual components as the project evolves.

Datasets

For this project, we will use the DS542 classroom Echo360 recordings as our primary dataset. This dataset consists of a series of lecture captures from a real-world classroom setting and includes multiple content streams, such as camera feeds, slides, and whiteboard annotations. By analyzing these recordings, we can extract relevant features and cues to inform our system's decision-making process. Additionally, we may consider augmenting this dataset with other publicly available lecture recordings to increase the diversity of our training data.

Evaluation

We will evaluate our system through both objective metrics and subjective assessments:

Objective Metrics

- Shot Selection Accuracy: We'll manually annotate optimal shot selections for a subset of lecture videos and compare our system's automated choices, targeting 85%+ agreement with expert annotations.

- Content Continuity: We'll measure the percentage of important teaching moments (e.g., equation derivations, key explanations) that are captured appropriately, aiming for 90%+ coverage.
- Technical Performance:
  - Real-time processing latency (target: $< 500ms$)
  - Frame rate stability (target: consistent $30fps$)
  - System resource utilization (target: $<$ 70% CPU/GPU on target hardware)

Comparative Evaluation

- Baseline Comparison: We'll establish the following baselines:
  - Static wide-angle recording (current standard)
  - Rule-based shot selection (e.g., switching to slides when detecting slide changes)
  - Human operator video direction (upper bound benchmark)
- Improvement Metrics: We'll measure the percentage improvement over the baseline static recording in terms of content visibility (ability to read text/see details) and instructor visibility.

Subjective Assessment

- Viewer Experience Surveys: We'll conduct surveys with students who view both our system-produced videos and baseline recordings, evaluating:
  - Content clarity and visibility
  - Engagement and attention maintenance
  - Overall learning experience quality
- Expert Evaluation: We'll have professional video producers and educational technology experts evaluate our system's output quality using standardized rubrics for educational video production.

Iterative Testing

- We'll use the ds542 lecture recordings as our development dataset, with a 70/30 train/test split.
- For final evaluation, we'll test on newly captured Echo360 recordings that the system hasn't seen during development.

- We'll conduct ablation studies to measure the contribution of each component (pose detection, speech understanding, etc.) to overall performance.

We expect to see at least a 40% improvement in viewer engagement and content visibility compared to static recordings, approaching the quality of professionally directed educational videos. The system should correctly identify and highlight at least 85% of pedagogically significant moments in a lecture. Through iterative refinement, we aim to reduce the gap between automated and human-directed educational videos while maintaining real-time performance.

Timeline

Week 1: Research and Planning

1. Conduct a literature review on related work
2. Define the system architecture and requirements
3. Set up the development environment and tools
4. Gather initial datasets

Weeks 2-3: Content Recognition

1. Implement OCR for lecture slides and chalkboard recognition
2. Develop a pipeline for transcribing spoken content
3. Evaluate the accuracy of content recognition on sample lecture recordings

Weeks 4-5: Gesture and Scene Analysis

- Develop gesture recognition models
- Person detection and pose estimation for instructor movements
- Evaluate gesture recognition accuracy on sample lecture recordings

Weeks 6-7: Video Composition

- Make the decision-making framework for the video composition
- Integrate multiple content sources
- Implement rules for switching between sources based on real-time scene understanding

Weeks 8-9: Integration and Testing

- Integrate the individual components into a unified system
- Connect it all to OBS or other streaming/recording software
- Test the system with real-world lecture recordings and refine as needed
- Conduct user testing and gather feedback

Weeks 10-11: Final Evaluation and Refinements

- Evaluate the system's performance against predefined metrics
- Gather feedback from users
- Prepare for project presentation

Conclusion

This project aims to develop an intelligent system for enhancing classroom video recordings by dynamically composing multiple content streams based on real-time cues. This will be accomplished by leveraging computer vision, automatic speech recognition, and content analysis. The system can create a more engaging, context-aware lecture capture that overcomes the limitations of a static camera feed. Through this project, we hope to maximize the potential of lecture capture systems by enriching the educational experience for remote learners.

References

[1] Panopto, "Panopto official website." [Online]. Available: https://www.panopto.com

[2] Echo360, "Lecture Capture Solutions." [Online]. Available: https://echo360.com

[3] Datavideo, "Easy and Simple Live Streaming for Lecture." [Online]. Available: https://www.datavideo.com/kr/solution/78/easy-and-simple-live-streaming-for-lecture

[4] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, 2013, pp. 6645–6649.

[5] R. Smith, "An overview of the Tesseract OCR engine," *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR)*, Barcelona, Spain, 2009, pp. 629–633.

[6] L.-P. Morency, "Modeling human communication dynamics [Signal Processing on Big Data]," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 112–116, Nov. 2010.