Smart Multimodal Classroom Video Recorder
Final Report

Bhavya Surana, Grace Chong, William Clavier

May 4, 2024

Abstract

We present a real-time lecture capture system that fuses camera feeds, slides, whiteboard text, and audio cues to dynamically compose context-aware video outputs. By integrating YOLOv8 for person detection, OpenPose for gesture recognition, Whisper for speech recognition, Tesseract for OCR, and a weighted decision engine, our pipeline selects and transitions among views to highlight pedagogical signals. The goal is to overcome the limitations of static cameras and provide a richer, more immersive experience for both live and recorded viewers.

1. Introduction

Static lecture recordings miss board annotations, slide references, and instructor gestures. Our system automatically detects teaching signals and applies rule-based fusion with weighted scoring (60% instructor activity, 40% content relevance) to select optimal views and overlays, enhancing engagement and accessibility without manual camera control.

2. Related Work

Traditional platforms (Panopto, Echo360) require manual cueing. Recent advances in computer vision (YOLOv8 [2], OpenPose [1]), ASR (Whisper [3]), and OCR engines (Tesseract [4]) enable automated composition. We build on these to create a modular real-time framework.

3. Datasets

Testing: A video maDS542 Echo360 recordings (10+ lectures, multi-camera, slides, audio). **Supplementary:** 50 hours of YouTube EDU lectures for diversity. **Annotations:** 200 slide frames, 100 whiteboard frames for OCR; 100 gesture cues.

3. Datasets

Testing Video: We produced a synthetic mock lecture video combining slide decks, scripted gestures (pointing, writing), and whiteboard annotations to validate pipeline performance under controlled conditions.

Annotations: 200 slide frames, 100 whiteboard frames for OCR; 100 gesture cues.

Supplementary: 50 hours of YouTube EDU lectures for diversity. DS542 Echo360 recordings (10+ lectures, multi-camera, slides, audio).

4. Methodology

4.1 Person and Pose Detection

Overview:
The pose estimation system is designed to detect and analyze the professor's pose in classroom videos, with a particular focus on identifying pointing gestures and writing cues. The system utilizes a deep learning-based approach using OpenPose architecture implemented through OpenCV's DNN module.

Technical Implementation:

1. Model Architecture:
    - The system employs the OpenPose architecture, specifically using the COCO model variant
    - The model is implemented using Caffe framework and consists of:
        - A prototxt file (pose_deploy_linevec.prototxt) defining the network architecture
        - A pre-trained model file (pose_iter_440000.caffemodel) containing the learned weights
2. Keypoint Detection
    - The system detects 18 key body points:

- Face points: Nose, Eyes, Ears
- Upper body: Neck, Shoulders, Elbows, Wrists
- Lower body: Hips, Knees, Ankles

- These points are connected through 17 predefined pairs to form the skeletal structure.

3. Pose Analysis Pipeline

- Preprocessing: Input frames are resized to 368x368 pixels. Pixel values are normalized to [0,1] range and Input is converted to a blob format suitable for DNN processing.
- Detection Process: The input frame is processed through the neural network. For each body part, the system:
  - Generates a probability map
  - Identifies the location with maximum probability
  - Applies a confidence threshold (0.1) to filter out low-confidence detections
- Pose Classification: The system analyzes the detected keypoints to determine:
  - Overall pose confidence based on the number of detected points
  - Arm angles using geometric calculations. Pointing gestures through specific criteria like Arm extension (angle $>$ 150 degrees) is classified as Pointing.

4.2 Slide-Speech Correlation

We implement a sophisticated correlation system that combines TF-IDF vectorization and cosine similarity to detect verbal slide references. The system:

- Vectorizes slide text and ASR output using TF-IDF with English stop word removal
- Computes cosine similarity between vectors to detect content alignment
- Applies temporal window analysis (5-second windows) to capture contextual relationships
- Uses confidence thresholds (0.3 minimum similarity) to determine slide relevance
- Incorporates zero-shot classification to enhance reference detection accuracy

A high similarity score ($>$0.7) triggers a view shift to the slide, while maintaining temporal coherence through window-based analysis.

4.3 OCR Pipeline

Approach: We apply SSIM-based frame differencing (SSIM [5]) to detect whiteboard changes, then crop regions for OCR. **Models:**

- OpenCV for frame loading and preprocessing.
- SSIM to detect content updates.
- Tesseract OCR for text extraction on slides.
- Custom timestamping aligns OCR outputs with gesture and transcription modules.

Limitations:

- Slideshow OCR struggles with diagrams and complex visuals.
- Attempted and decided to remove: Whiteboard OCR accuracy is low (54% raw Tesseract, 68% Vision API) due to handwriting variability and camera resolution.

4.4 Visual Overlay Safety Analysis

Our overlay system employs a multi-stage approach to ensure safe text placement:

- Edge detection to identify content-dense regions
- Corner scoring based on content density and edge proximity
- Adaptive confidence thresholds for corner safety assessment
- Temporal consistency checks to prevent rapid overlay shifts

This ensures overlays do not obscure key visuals while maintaining readability.

4.5 Fusion and Decision Engine

We compute a weighted decision score that integrates multiple modalities:

- 60% from instructor activity (gesture detection and pose analysis)
- 40% from content relevance (slide-speech correlation and OCR)
- Confidence adjustments:
  - 1.5x boost when modalities agree (e.g., pointing gesture + slide reference)
  - 0.7x penalty on conflicts (e.g., board writing + slide reference)
- Temporal smoothing to prevent rapid view switching

The system also incorporates deictic reference analysis to enhance context awareness, detecting phrases like "look at this" or "as shown here" to improve view selection accuracy.

5. Evaluation

Our evaluation metrics were obtained using a local evaluation harness processing manually labeled datasets stored alongside the pipeline code. Results:

- **Slide OCR:** 92% character-level and 85% word-level accuracy on 200 manually transcribed slide frames.
- **Whiteboard OCR:** 68% word-level accuracy (vision-enhanced) vs. 54% raw Tesseract on 100 annotated board crops.
- **Gesture Classification:** 78% accuracy distinguishing pointing versus walking over 100 labeled frames.
- **Cue Detection:** 80% precision and 77% recall against 100 annotated key moments.
- **Reference Detection:** 85% accuracy in identifying slide and board references using zero-shot classification.
- **Overlay Safety:** 82% of overlays correctly placed in safe regions based on content density analysis.
- **Latency:** End-to-end decision under 200ms per frame on an NVIDIA RTX 2080 GPU.

6. Technical Choices & Insights

OpenPose: Chosen for robust multi-person accuracy and pretrained models.

Tesseract OCR: Lightweight, easily integrated.

Weighted Fusion: Balances activity cues and content relevance for reliable composition.

7. Challenges

- *Pointing Ambiguity:* Side angles obscure skeletal keypoints, causing false positives.
- *Lighting Variability:* Shadows during board annotations can trigger spurious gestures.
- *Handwriting Variability:* Loose, cursive writing reduces OCR recall.

- *Decision Making:* Making a decision that makes visual sense for what to show on the screen is harder than it would appear.

8. Conclusion

Our modular, real-time system effectively composes lecture videos by fusing vision, text, and audio. Future work includes integrating a transformer-based context fusion network and exploring reinforcement learning for smoother, adaptive shot selection.

References

[1] Z. Cao et al., "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," CVPR, 2017.

[2] G. Jocher et al., "YOLOv8: State-of-the-Art Real-Time Object Detection," GitHub, 2023.

[3] A. Radford et al., "Whisper: Robust Speech Recognition via Large-Scale Weak Supervision," OpenAI, 2023.

[4] R. Smith, "An overview of the Tesseract OCR engine," ICDAR, 2009.

[5] Z. Wang et al., "Image Quality Assessment: From Error Visibility to Structural Similarity," IEEE Trans.