OPEN ACCESS

International Journal of Innovations in Science & Technology

Deep Learning-Based Automated Classroom Slide Extraction

Zeeshan Azhar[1], Hassan Nazeer Chaudhry[1], Farzana Kulsoom[2], Sanam Narejo[3]

[1]Computer Science Department, Barani Institute of Technology, Rawalpindi, Pakistan

[2]Telecommunication Engineering Department University of Engineering and Technology Taxila, Pakistan.

[3]Department of Computer Systems Engineering. Mehran University of Engineering and Technology, Jamshoro.

* **Correspondence**: Hassan Nazeer Chaudhry Hassan@biit.edu.pk

Citation | Azhar. Z, Chaudhary. H. Z, Kulsoom. F, Narejo. S, "Deep Learning-Based Automated Classroom Slide Extraction", IJIST, Vol. 6 Issue. 2 pp 380-395, April 2024

Received | Mar 22, 2024, **Revised |** April 17, 2024, **Accepted |** April 20, 2024, **Published |** April 25, 2024.

Automated extraction of valuable content from real-time classroom lectures holds significant potential for enhancing educational accessibility and efficiency. However, capturing the spontaneous insights of live lectures often proves challenging due to rapid visual transitions, instructor movement, and diverse learning styles. This paper presents a novel approach that combines the strengths of YOLO and Scale-Invariant Feature Transform (SIFT) techniques to automatically extract slides from live classroom lectures. YOLO, a real-time object detection algorithm, is employed to identify board area, teacher, and other objects within the video stream. While SIFT, a robust feature-based method, was used to accurately merge key points from multiple pictures of the same region. The proposed method involves a multi-stage process: first, YOLO detects the potential place of the teacher, which occluded the board within the video frames. Subsequently, the teacher was removed from the image. The board was divided into multiple segments, to remove and merge redundant content Scale-invariant feature Transform (SIFT) was employed. Experimental results on a diverse dataset of classroom lecture videos demonstrated the effectiveness of the proposed method in extracting slides across different environments, lecture styles, and recording conditions. The potential benefits include improved note-taking, reduced manual effort in content curation, and enhanced accessibility to lecture materials. The presented approach contributes to the broader goal of leveraging computer vision and machine learning techniques to transform traditional classroom settings into modern, interactive, and adaptive learning environments.

Keywords. Deep Learning; Computer vision; Academic assistant system; YOLO; Object detection; CNN.

Introduction:

Capturing valuable knowledge from live classroom lectures presents a persistent challenge for students and educators [1]. The dynamic nature of lectures, characterized by rapid visual transitions, instructor movement, and diverse learning styles, can lead to information overload and difficulty in retaining key points. Traditional methods including manual note-taking or audio recording often require significant effort, are prone to errors, and lack the flexibility to cater to different learning preferences. Leveraging computer vision for automated content holds significant potential for revolutionizing various fields, particularly education and training. Similarly, Educational Assistant Systems (EAS) utilize information technology to simplify classroom responsibilities for teachers, for example, student performance monitoring, grading, and assessments, and improve students' learning experience. Thus, EAS can significantly reduce teachers' workload and allow them to focus on more strategic and interactive aspects of teaching, ultimately leading to a more engaging and effective learning experience for students. This technology utilizes computer vision, deep learning, and data analytics to extract valuable insights from educational settings. These computer vision technologies are used in several areas such as monitoring student behavior and engagement [2][3][4], automating e-boards [5][6], extracting lecture slides [7], detecting cheating in exams [8], and other fields [9]. Object detection and feature extraction are essential techniques in computer vision that play a crucial role in various applications such as object recognition, tracking, and image analysis [10].

In the context of educational challenges, institutions grapple with numerous issues in a typical classroom-learning environment daily. As the role of educational tools and resources expands on the Internet, the prevalence of video recordings capturing classroom lectures is growing. In support of their students, educators frequently upload these recorded lecture videos online. These videos serve as supplementary aids, integral components of coursework, and even substitutes for traditional in-person classroom sessions, as seen in the flipped classroom model. Notably, a considerable number of instructors distribute these videos to a broader online audience, leading to a growing demand for software solutions that help students navigate extensive collections of lecture recordings effectively [11][12]. Such software would seamlessly integrate note-taking and review processes into a digital environment optimized for video viewing. By providing students access to these lecture videos, substantial benefits emerge. These videos empower students with the ability to control the pace at which they engage with the educational material, serving as valuable study resources that enhance the overall learning experience.

The study acknowledges the significant benefits lecture video series offer to students while also highlighting the challenges, they pose in terms of navigation and the inability to fully replace provided notes, slides, or student-generated materials. The lack of a well-defined structure makes it challenging to point out specific objects within the lecture, titles and descriptions of videos often fall short representing the breadth of topics covered. Additionally, visual content integral to the learning experience is often lacking. Students encounter limited options when it comes to effectively navigating through such videos, as conventional video scrubbing using cursor-driven visual search, proves to be a slow and unreliable method for pinpointing. A potential solution to this challenge, involves segmenting the video according to the lecturer's slides and it could also be applied to PowerPoint-based lecture video recordings, this would allow students to select a slide they want to review, prompting the system to automatically navigate the corresponding segment of the recording, which presents the specific slide. The synchronization or extraction of slides from videos constitutes its own set of challenges, as acknowledged in previous research [2][3]. However, this study delves further into the complexity of synchronizing and extracting slides from lecture videos involving chalkboard presentations, wherein the lecturer intermittently obscures distinct portions of the board at varying times, and the points of transition lack distinct clarity or suddenness.

In general, the methodologies commonly employed for extracting and segmenting slides in slideshow presentation videos cannot be directly transposed to scenarios involving lecturers employing visual aids like a chalkboard, a prevalent occurrence in STEM courses. The distinctions arise primarily due to the absence of pre-existing electronic slides for video segmentation. Furthermore, a significant portion of the visual content within these videos stems from the lecturer's dynamic interactions, taking place prominently in the foreground, often diverging from the actual lecture content. An added complication emerges from the lecturer frequently obstructing portions of the chalkboard, a situation seldom encountered in presentations structured around slides. The techniques expounded in this paper serve to streamline video navigation through video segmentation. In the context of a chalkboard-based lecture, the methodologies delineated within this study empower the system to eliminate obstructive elements from the learner's perspective of the chalkboard within the lecture recording. Additionally, these methods facilitate the division of the lecture into distinct slides, effectively enhancing the comprehensibility of the video content.

The system introduced in this context alleviates students from the laborious aspects of note-taking, such as the meticulous transcription of chalkboard content or the synchronization of slides and notes with video in cases involving slideshow presentations. This allows for a more engaged and interactive note-taking experience that encourages deeper engagement with the material, all while requiring no additional effort from educators. Additionally, this system is versatile enough to be retroactively applied to videos that were not originally recorded with previous approaches. Within this paper, we outlined a dual-layer framework for chalkboard-based lecture videos.

Our devised system takes a standard chalkboard lecture video and produces two distinct outputs: a summarised slideshow capturing the lecture content, and a video notepad, a condensed video with the lecturer's presence removed where students can input their notes. The algorithm is responsible for extracting the video notepad operates in real-time, enabling students to annotate content while the lecture unfolds. Subsequently, we further refine this abridged, lecturer-free video notepad to generate a minimal series of static images, or" slides," captured before any content is erased from the chalkboard. This set of slides serves as a concise overview of the content within each video, offering a swift means to search numerous videos for specific slides. Moreover, these slides can function as study aids or references, akin to traditional lecture notes. This hierarchical representation facilitates rapid navigation through extensive video lecture collections, promotes effective note-taking that encourages generative learning, and also serves as a reference for video content, analogous to a slideshow. Empirical results validating the proposed segmentation algorithm are also presented.

One of the primary challenges in education is ensuring that students receive accurate and comprehensive lecture notes. In situations where teachers write on boards, students often manually transcribe lecture notes, leading to errors during the writing process. Moreover, the teacher may have to take occasional pauses to allow students to synchronize their notes, resulting in a loss of valuable teaching time [13]. Furthermore, manual note-taking is susceptible to omissions, where students may miss crucial lecture points [14]. While teachers may provide supplementary reading materials, they often share additional information during lectures, making manual note-taking essential.

The manual preparation of lecture materials, specifically the extraction and organization of lecture slides, poses a time-consuming and repetitive task for educators. As educators strive to deliver engaging and impactful lectures, the overhead of creating and refining slide content can divert their focus from instructional design and content delivery intricacies. Similarly, students face the challenge of effectively assimilating and reviewing lecture content, often resorting to labor-intensive note-taking and managing disparate materials. These inefficiencies hinder the realization of the full potential of education, where educators and learners can engage in meaningful interactions that foster deep understanding and critical thinking. Consequently, Human activity recognition [15][16] is well well-established task in computer vision and can be employed for improving the learning experience of students in the classroom.

Acquiring slides during a lecture using computer vision presents several challenges that need to be addressed. Firstly, the teacher often obstructs the board when writing on it, which makes capturing clear images a challenging task. Additionally, multiple frames of the board need to be captured and carefully stitched together to obtain unified data, considering that there might be redundant information across different boards. Moreover, the real-time responsive nature of the lecture needs computation to be fast enough to detect the teacher, maintain key points of the board, and generate slides.

Objectives:

The objective of this study is to enhance the accessibility, review ability, and annotatability of lecture videos for students, all the while preserving the flexibility of lecturers' teaching methods. Specifically, this research focuses on optimizing lecture videos that involve the manual drawing of figures on a chalkboard (or whiteboard). This research offers the following key contributions:

- We provided an audio synchronization with slides where real-time lecture room voice is recorded and played at the right positions as an audio transcription of slides.
- We developed a novel algorithm that effectively stitches together multiple frames of the chalkboard at various writing points, creating a cohesive and obstruction-free slide.
- The system keeps track of the progress of the lecture over time and arranges lecture slides into consolidated lecture slides synchronized with audio.

Our research introduces a method to identify and rectify areas that were inadvertently removed or obscured during the lecture due to certain mistakes. This ensures that the final lecture material is comprehensive and accurately represents the content delivered during the lecture. To address obstructions caused by the teacher at different points in the lecture, our system intelligently fills in these occluded areas by referencing unobstructed portions from other segments of the lecture.

In recent years, object detection has experienced a revolution due to deep learning approaches, notably convolutional neural networks. Deep learning methods include the well-known You Only Look Once (YOLO) algorithm, which has outperformed in terms of accuracy and speed [17][18]. Another popular method for feature extraction is Scale-Invariant Feature Transform (SIFT). In computer vision, SIFT is a commonly utilized feature extraction approach. The real-time performance of the YOLO (You Only Look Once) algorithm stands out as one of its primary advantages. This algorithm's ability to rapidly detect objects makes it well-suited for applications that require real-time operation, especially in scenarios where prompt decision-making is essential [19]. Recently, the utility of technology in education has increased notably, especially in applying deep learning methods to automate different educational tasks. One intriguing area of study involves the automated retrieval of classroom slides from image-based lecture capture systems.

The authors in [20] introduced a method for automatically generating presentation slides from spoken content. In order to extract spoken content and condense important insights into visually appealing slides, their methodology used effective voice recognition and natural language processing techniques. The authors sought to improve the efficiency of content development without sacrificing the purity and formativeness of the slides by concentrating on academic presentations. Additionally, [21] summarizes the challenge of summarizing extensive lecture videos into concise and informative slides [21]. The researchers employed sophisticated video analysis techniques in order to identify slide transitions and automatically extract pertinent content. By employing text summarisation methods, the authors aimed to generate concise transparencies that encapsulated the fundamental ideas deliberated in the lecture, thereby assisting students in comprehending and reviewing the material more easily. The researcher in

[22] made a significant contribution by creating an automated system to index lecture videos. Her approach allowed easy navigation in educational videos by detecting slide content by processing images and extracting text with Optical Character Recognition (OCR). Using this indexing system, both teachers and students could quickly find the exact video clips that matched the slides used in class. Researchers in [23] presented a plan for improved semantic understanding of educational videos. It extends the FitVid dataset to suit any learning video and enhances existing slide classification techniques. The goal is to detect "slide" or "non-slide" shots using heuristics like size, location, numbers, and object types.

The study in [24] initially converts the segmented lecture videos into keyframes for minimal duplication, extracting textual information, and addressing text detection and recognition. Moreover, a binarization approach optimizes text locations, VietOCR recognizes English and Vietnamese text, and a vector-based semantic search in Elastic Search enhances search performance during a video lecture. Experimental results show high performance. The authors in [25] presented a novel procedure for teachers using traditional formats like slideshows to enhance their education towards open education. It improves findability by breaking down teaching material into fragments and enhancing them with metadata. This approach enhances accessibility through open internet access, ensuring that the process remains future-proof. The system is based on industry-standard datatypes and community-driven representations and is reusable through the editable and extensible nature of output H5P containers. The system can be easily extended to handwritten text on a digital blackboard. The prototype focuses on slide decks and links background information within the teacher's toolkit. The study in [21] proposes leveraging the transfer learning approach to explore the portability of pre-trained video meeting summarization models to abstract video lectures, overcoming the challenge of training ad hoc deep learning models on large human-annotated datasets. It aims to enhance the accessibility of video lecture content for learners with special needs and challenges. Furthermore, it produces a synthetic textual summary of key concepts, generating more readable and actionable summaries than previous methods. The method also reuses pre-trained models for meeting summarisation. However, the experimental results indicate that it generates more fluent and actionable summaries than content extraction. The continuous refinement of methodologies and frameworks promises to significantly impact pedagogical practices, enhancing the efficiency, accessibility, and effectiveness of educational materials.

Material and Methods.

Our paper introduces a novel method for automatically extracting slides from live classroom lectures, addressing challenges such as rapid visual transitions and diverse learning styles. By combining YOLO and SIFT techniques, our approach efficiently identifies key elements within the lecture video stream. YOLO is utilized for real-time object detection, identifying the board area, teacher, and other objects, while SIFT accurately merges key points from multiple images of the same region. YOLO's capacity to identify several things in a single pass is one of its other strong areas. This indicates that the YOLO algorithm produces accurate and efficient results by concurrently detecting and classifying several objects in an image. YOLO is particularly useful for generalizing from natural photos to other domains e.g., artwork, etc. because it can learn general representations of objects. Additionally, unlike earlier techniques like R-CNN, YOLO is a one-stage object identification algorithm, implying it does not require complex optimization or multiple steps. The detailed flow of our research is presented in Figure 1. In the initial step of our research, we captured high-definition frames from a live video feed along with corresponding timestamps. Simultaneously, audio recording is initiated to ensure precise synchronization between visual and auditory components as referred to in [26]. This synchronized multimedia data serves as the foundation for the processing stages. The basis for processing steps is the synced multimedia data. Leveraging the YOLO object detection algorithm, which effectively locates and detects important components of the environment of the class, notably the person and the lecture board. For additional content extraction and analysis, this stage serves as the foundation.

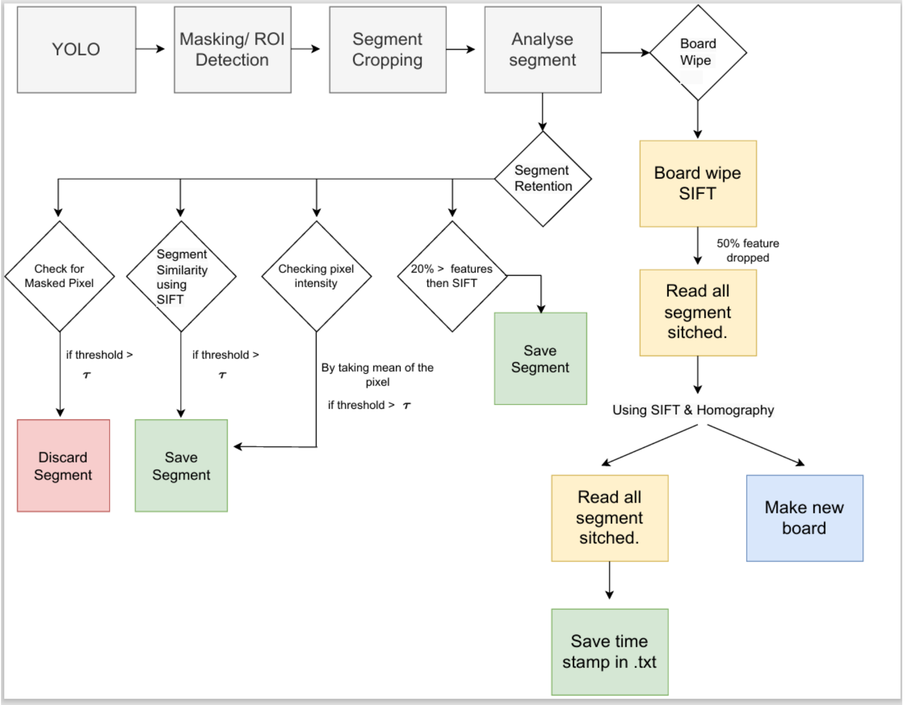

Figure 1. Workflow of Proposed Research Methodology.

Data Collection and Acquisition:

The dataset was developed at the Barani Institute of Information Technology, Rawalpindi, Pakistan with recordings done over three months between September 2023 to January 2024, in the fall semester of 2023. At the Barani Institute of Information Technology, creating the dataset capturing classroom scenarios required extensive planning, meticulous execution, and continuous refinement. The goal was to develop a resource that not only reflected the diversity of educational settings but also provided a rich and realistic environment for training and testing algorithms in educational technology.

Each image in the dataset had a resolution of 608x608 pixels. This resolution provided a balance between capturing sufficient detail and maintaining manageable file sizes. The images depicted classroom scenes with both teachers and students present. As mentioned in the earlier section this reflects the real-world environment where the system is intended to function. The data set was further divided into (Train-Validate-Test) split having 70% for training, 20% for validation, and 10% for testing. A number of YOLO configurations were considered to find a balance between speed and accuracy.

The study was carried out with a limited set of images consisting of 263 images that depicted various classroom situations. During training, the models learn to identify relevant objects and features such as the board and slide content within the images. However, the validation set was used to fine-tune the models during the training process. The model's performance was evaluated on the validation set, and hyperparameter settings that control the model's behavior can be adjusted to improve accuracy and avoid overfitting. Figure 2 presents the cumulative visualization of training images showing the classroom environment and objects. The images were annotated using the online RoboFlow toolchain to obtain the ground truth for objects to be recognized.

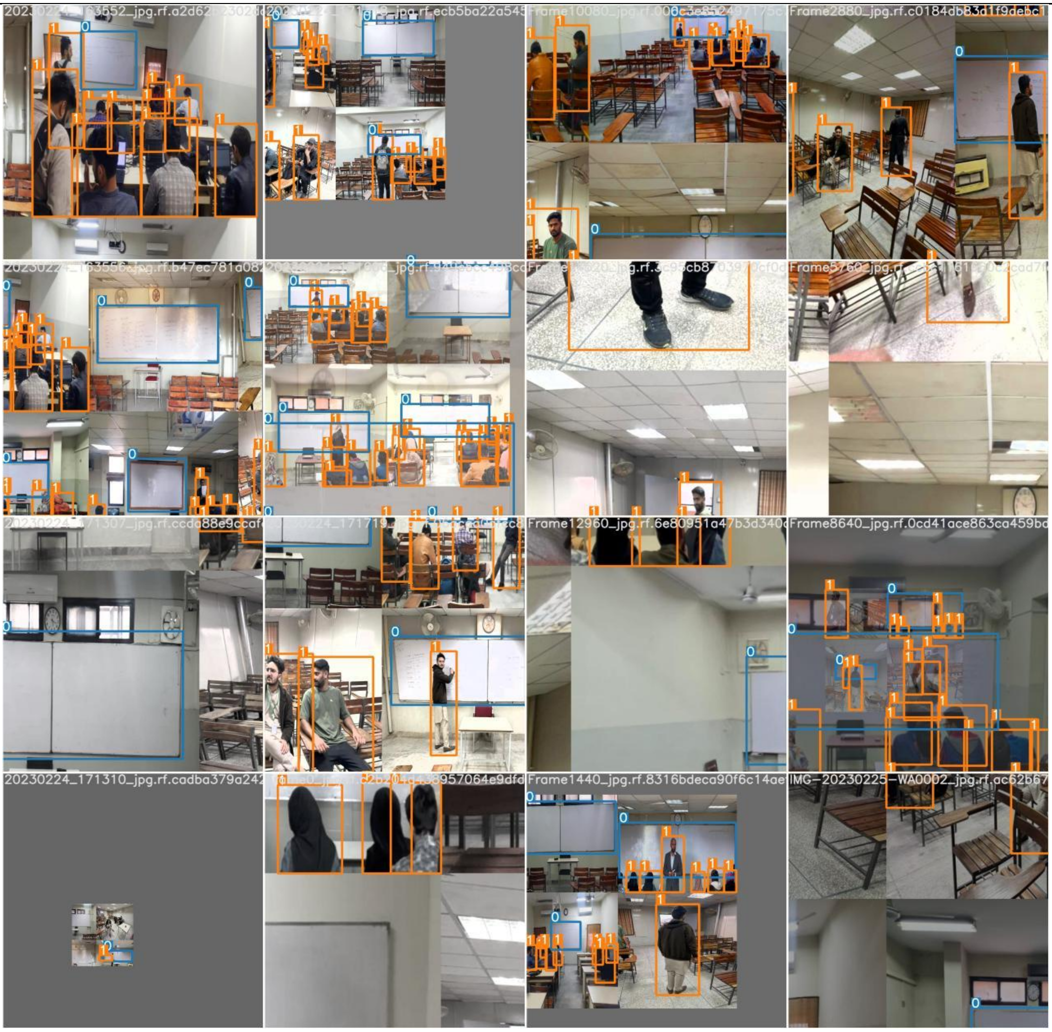

Figure 2. The cumulative visualization of training images showing the classroom environment and objects.

Experimental Setup:

We conducted our experiments using NVIDIA GPU RTX 3070, with 128 GB of RAM and Intel core i7 12th generation, ensuring efficient processing for our tasks. This system was implemented using Python, utilizing the power of the PyTorch framework with an emphasis on YOLO v7 architecture. Our algorithm performance and speed were improved by PyTorch CUDA acceleration, which speeds up both board and person detection.

Dataset Collection and Challenges Faced in Real-World Classroom Scenarios:

Recognizing that classrooms vary widely in terms of layout, furniture arrangement, and student demographics, efforts were made to capture a broad spectrum of scenarios. Classrooms of different sizes, configurations, and levels of technological integration were included to ensure the dataset's relevance across various educational contexts. Furthermore, scenes were captured across various subjects, ranging from mathematics and science to humanities and arts. This diversity was crucial for ensuring that the dataset could be used to develop algorithms applicable to a wide range of educational disciplines. However, as with any real-world data collection effort, challenges inevitably arose during the process. One such challenge was the presence of students occupying front rows, partially obstructing the view of the classroom scene. While this posed a logistical challenge for capturing clear images, it also presented an opportunity to incorporate realistic occlusion scenarios into the dataset. By including images where students partially obscured the view, we could train algorithms to accurately detect and interpret such occlusions, a crucial capability for real-world applications where visual obstructions are common. To address this challenge, careful documentation was essential. Each image was annotated with detailed metadata, including information about the classroom layout, the subjects being taught, and the positioning of teachers and students. This metadata not only provided valuable context for understanding each image but also facilitated the development of algorithms for scene understanding and analysis. Moreover, the dataset was continually refined and augmented over time. As new scenarios and classroom settings were encountered, additional images were captured to enrich the dataset further. This iterative process ensured that the dataset remained relevant and up-to-date, reflecting the evolving landscape of educational technology and classroom practices.

Teacher and object detection in Class:

Several occlusions, such as students' heads in the front row and objects like tables, obstruct the view of the board and teacher. The YOLO is a well-known technique used for object detection in real-time as mentioned in earlier sections. Its introduction in 2015 marked a significant departure from previous methods like Sliding Window object identification, RCNN, Fast R-CNN, and Faster R-CNN. While processing an image, it not only detects objects but also delineates their spatial characteristics. This includes assigning a probability of a class (Pc), denoting the confidence level of object presence, as well as determining the object's center coordinates (Bx, By), bounding box dimensions (Bw, Bh), and classification (C1, C2, ... Cn). One of YOLO's standout features is its ability to handle multiple object detections seamlessly. Traditional neural networks often struggled with scenarios where multiple objects coexist, as they were primarily designed for single-object outputs. YOLO overcomes this limitation through a grid-based strategy, dividing the image into smaller cells. This grid analysis produces a comprehensive 7-size vector output that contains crucial data such as object presence, spatial coordinates, and classification. YOLO has various limitations, one of the challenges includes more localization errors compared to other state-of-the-art detection systems. Other challenges include difficulty in detecting objects at smaller scales or distant targets, as the shooting distance can introduce large variations in scale for the targets. Additionally, it may struggle with detecting objects in complex scenes, for example, the lightening variations of extreme darkness or extreme brightness or uneven lightness and cluttered backgrounds such as Busy backgrounds or multiple boards.

However, a strong resolution is required due to the complexity of overlapping bounding boxes. To address this issue, YOLO makes use of the Intersection over Union (IOU), which aids in accurately delineating object boundaries, especially in complex scenarios with overlapping objects. Consequently, redundant boxes were wisely suppressed, improving the accuracy of object detection results. YOLO has clearly changed the standards for real-time object identification by using equations spanning Pc, Bx, By, Bw, Bh, C1, C2, and Cn*:* as expressed below:

$$P_c = \sigma(t_{conf})$$
$$B_x = \sigma(t_x) + c_x$$
$$B_y = \sigma(t_y) + c_y$$
$$B_w = p_w \cdot e_{tw}$$
$$B_h = p_h \cdot e_{th}$$

Where the variable Pc represents the probability of an object's presence within a given bounding box. This probability is determined by applying the sigmoid function to the predicted confidence score t_{conf}. The variable Bx corresponds to the x-coordinate of the center of the bounding box. It was calculated by adding the result of the sigmoid function applied to the predicted t_x to the offset. Similarly, the variable represents the y-coordinate of the bounding box's center. This value was computed by adding the sigmoid function of the predicted t_y to the c_y offset. The term Bw denotes the width of the bounding box. It was obtained by multiplying the anchor box width p_w with the exponential of the predicted t_w. Lastly, B_h indicates the height of the bounding box. This value was derived by multiplying the anchor box height p_h with the exponential of the predicted t_h. In essence, these equations describe how YOLO's neural network architecture calculates and refines the bounding boxes' characteristics, including position, size, and object probability, thereby enabling accurate object detection. The IOU is calculated mathematically using the equation given below:

$$\text{IOU} = \text{Area of Intersection}/\ \text{Area of Union}$$

Where:

$$\text{Area of Intersection} = \text{Width of Overlap} \times \text{Height of Overlap}$$

$$\text{Area of Union} = \text{Area of Bounding Boxes} - \text{Area of Intersection}$$

$$\text{Pr (classi)} = \text{Pr (object) IOU truth pred} \cdot \text{Pr (classi} \mid \text{object)}$$

Where:

- Pr (object): Probability of object presence in the bounding box.
- IOU truth pred: Intersection over Union between predicted and ground truth boxes.
- Pr (class I | object): Conditional probability of class I given the presence of an object.

Classification and localization were seamlessly integrated into a unified process within the YOLO algorithm. YOLO manages numerous object detections and its strategic use of IOU. Following successful object detection, the YOLO algorithm masks the person's position with distinct pixels. The boundaries of the presentation board were precisely extracted through cropping techniques. This step made sure that the processing afterward focused only on what was needed within the board region. The cropped image was separated into six separate portions to ensure a thorough investigation of the presentation board's information. The thorough examination of each image portion was primarily meant to confirm the person's presence and

Segment Acceptance Criteria.

Specific pixels inserted by the YOLO object detection algorithm were examined in this research to ascertain the person's presence on the board, indicating their location within the board area. The number of designated pixels within each part of the board was counted and examined in detail. We determined if the presenter's presence significantly blocks the board's content in that part by comparing this designated pixel count against a predetermined threshold. The presenter significantly covered the board if the designated pixel count was higher than the threshold. The segment was flagged as obscured and was excluded from further examination and processing. The SIFT and ORB feature detectors were used to assess each section's quality and distinctiveness in addition to occlusion detection. These detectors help in the evaluation of content changes and adjustments by offering insights into the distinctiveness and qualities of the visual elements inside the section.

To handle situations where a substantial portion of the board's content was erased or altered, we considered initiating a board wipe if the current segment retains less than 50% of the features from the previous segment. This initiates an image-stitching process that seamlessly joins all image segments together, assuring consistency and completeness in the visualization of the board. A variety of parameters are utilized for making decisions regarding segment retention, including assessments of visual clarity and quality made by intensity and sharpness analyses. SIFT was used to measure picture similarity, and a specified dissimilarity threshold (0.1%) determines which segments are accepted. An additional factor in decision-making is the comparison of feature counts with the prior section, with a 20% rise indicating the potential addition of further content, perhaps in the form of additional written material.

In order to recreate the lecture board, all image segments were smoothly stitched together, preserving a complete and clear image of the board along with timestamps of board construction. This procedure makes sure that the lecture's progress on the board is accurately recorded. The entire process described above runs iterative during the lecture. The dynamic nature of the technology ensures that changes to the content of the board are continuously tracked and recorded, leading to a dynamic visual representation of the lecture. Upon lecture completion, the recorded board images and their start and end times for the audio recording were sent to the slide generation module. This module separated the audio recording into separate clips that belong to each distinct board slide. These voice recordings were then combined with the appropriate board images, which resulted in the creation of a complete PowerPoint (.pptx) presentation.

In this study, we employed Algorithm 1 which summarises the overall flow of our work, the input to our system is synchronized multimedia data from a video feed, lecture environment, and audio recording. It generated processed lecture content with synchronized audio in a .pptx file. The first step involves data Capture and synchronization, which captures high-definition frames with time stamps from live video feed. It also initiated audio recording concurrently for precise synchronization. The second step in algorithm 1 is object detection and extraction which initializes the YOLO algorithm. Resulting in the detection and localization of lecture components such as (person, and lecture board). Finally, regions of interest for content analysis were extracted, following pixel masking, and cropping was done by applying pixel masking to the person's position using YOLO results.

We also employed cropping techniques to isolate the boundaries of the presentation board. In the next step, thorough examination and segmentation were performed by dividing the cropped image into six segments for detailed analysis, evaluating each segment for the presence of key elements. The fifth step is person presence determination by analyzing designated pixels from YOLO to ascertain a person's presence. Moreover, pixel count was calculated and compared against a threshold. Afterwards, a quality assessment is conducted. Subsequently, the content adjustment utilizing SIFT and ORB detectors for feature analysis is done. Similarly, determining segment acceptance based on visual clarity and feature count. It could be seen in Algorithm 1 that the seventh step is adaptive segment retention which identifies potential board wipes if the current segment has < 50% features of the prior segment, and finally stitch segments to maintain content consistency. This follows board reconstruction and visualisation seamlessly merging image segments for board reconstruction. Also, it is important to preserve the reconstructed image with timestamps of board construction. As the last step slide generation and presentation Creation is employed to compile recorded board images and audio timestamps. Separate audio into distinct board slide clips to generate PowerPoint presentations by combining audio and board images. It should be ensured that dynamic tracking and continuous recording continuously track changes in content and lecture progress.

Algorithm 1: Methodology and Proposed Technique

```
Input: High-definition frames, timestamps, YOLO results
Output: Processed lecture content with synchronized audio in .pptx file
procedure PROCESSLECTURECONTENT
image ← Capture Image and Timestamp () {Capture live video frame with timestamp}
processed Image ← YOLO Object Detection (image) {Locate and detect components}
masked Image ← Apply Pixel Masking (processed Image) {Mask person's position}
cropped Images ← Crop Image (masked Image) {Extract board boundaries}
image Array ← {} {Array to store accepted images}
for all segments in cropped Images do
   person Present ← Check Person Presence(segment) {Check masked pixel count}
   if person P resents then
     Discard segment {Person significantly blocks content} else
     Keep segment {Segment passes acceptance criteria}
     image Array ← image Array ∪ {segment}
   end if
   board Wipe ← Check for Board Wipe (segment) {Check for 50% content loss}
   if the board Wipe then
     Stitch Images (image Array) {Join segments for content consistency}
     break
```

```
        end if
        content Change ← Check for Content Change (segment, previous Segment) {Check for 20% content increase}
        if content Changes then
            Process Image Adjustment (segment) {Analyze quality and distinctiveness}
        end if
    end for
    reconstructed Board ← Merge Image Segments (image Array) {Reconstruct lecture board}
    Generate Dynamic Visualization (reconstructed Board) {Continuous tracking and recording}
    Audio Clips ← Separate Audio Recording () {Split audio recording into clips}
    Generate PowerPoint Presentation (reconstructed Board, audio Clips) {Create .pptx presentation}
    end procedure=0
```

Result and Discussion:

This section presents the results of our automated lecture slide creation system along with audio. The experiment was conducted on a small dataset.

Board and Person Detection:

The core component of the system is its board and person detection module, crucial for locating and classifying board and person, thereby playing a significant role in identifying and segmenting the board for lecture slide creation. Leveraging the YOLO v7 architecture in the PyTorch framework, the model was fine-tuned on a specialized dataset, resulting in an impressive accuracy of 96% during evaluation.

Figure 3 showcases the loss and accuracy of the trained model, highlighting its performance metrics. Additionally, Figure 4 displays the benchmarks of the trained YOLO v7 model, providing further insights into its capabilities. Table 1 summarizes evaluation metrics for various classes in the training dataset, offering a comprehensive overview of the model's performance across different categories.

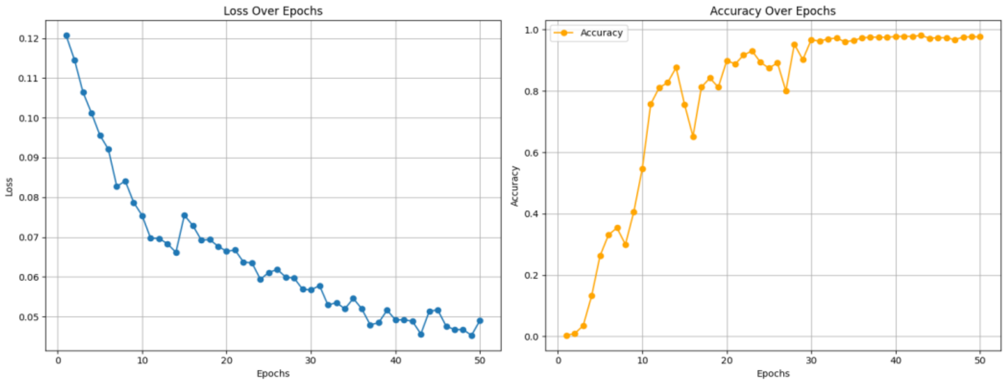

Figure 3. The performance of the trained model in terms of Loss and Accuracy

The analysis of the model's performance reveals significant insights. In Figure 3, it is observed that after 40 epochs, the loss was decreased notably, accompanied by an accuracy of 92%. This trend was similarly reflected in the mapping score depicted in Figure 4. It could be seen from Evaluation Metrics for various classes in Table 1 that the mAP score for detection of the board provided the highest value of 0.99 while for the person it was 0.95. It could be noted that the scores for mAP@.5 calculated the average precision when the IoU (Intersection over Union) threshold was set to 0.5. While mAP@.5:.95 calculated the average precision by averaging the precision values across different IOU thresholds, ranging from 0.5 to 0.95.

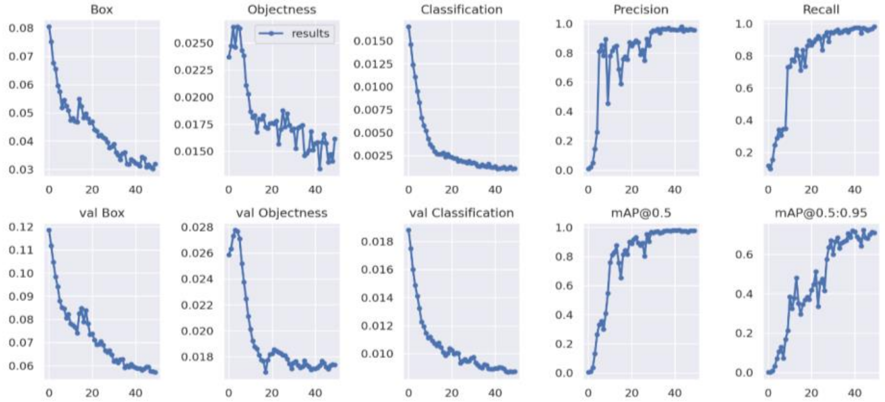

Figure 4. YOLOv7 model's benchmarks.

Table 1. Evaluation Metrics for various classes

Class	Images	Labels	P-score	R-score	mAP@.5	mAP@.5:.95
all	54	253	0.956	0.981	0.977	0.71
board	54	39	1	0.999	0.995	0.842
person	54	214	0.911	0.963	0.959	0.579

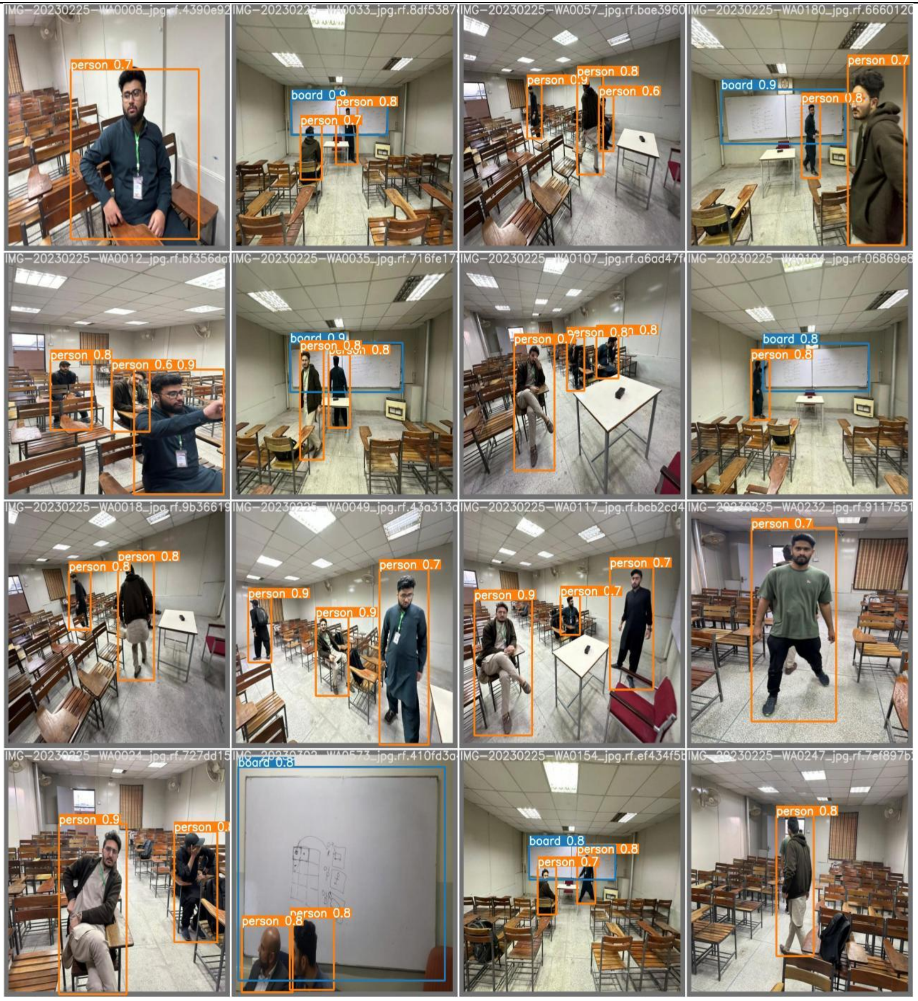
IMG-20230225-WA0008_jpg.rf.4390e92
person 0.7
IMG-20230225-WA0033_jpg.rf.8df5387
board 0.9
person 0.8
person 0.7
IMG-20230225-WA0057_jpg.rf.bae3960
person 0.8
person 0.9
person 0.6
IMG-20230225-WA0180_jpg.rf.6660120
board 0.9
person 0.7
person 0.8
IMG-20230225-WA0012_jpg.rf.bf356da
person 0.8
person 0.6 0.9
IMG-20230225-WA0035_jpg.rf.716fe17
board 0.9
person 0.8
person 0.8
IMG-20230225-WA0107_jpg.rf.a6ad47f
person 0.7
person 0.8
IMG-20230225-WA0104_jpg.rf.06869e8
board 0.8
person 0.8
IMG-20230225-WA0018_jpg.rf.9b36619
person 0.8
person 0.8
IMG-20230225-WA0049_jpg.rf.43a313
person 0.7
person 0.9
person 0.9
IMG-20230225-WA0117_jpg.rf.bcb2cd4
person 0.7
person 0.9
person 0.7
IMG-20230225-WA0232_jpg.rf.9117551
person 0.7
IMG-20230225-WA0024_jpg.rf.727dd15
person 0.9
person 0.
board 0.8
person 0.8
person 0.8
IMG-20230225-WA0154_jpg.rf.ef434f5
board 0.8
person 0.8
person 0.7
IMG-20230225-WA0247_jpg.rf.7ef897b
person 0.8

Figure 5. Test images showing the object detection of board and students as a person

Automated Lecture Slide Creation:

This module automated the lecture slide creation process, we tested this module by imitating real-world classroom situations where teachers wrote on the board while students were present. The results, as depicted in Figure 5, illustrate the algorithm's ability to generate slides for lectures using segmentation and stitching techniques. The results showcased an accuracy of 92%, demonstrating its ability to capture and deliver board content without obstructions.

Audio Synchronisation:

The seamless integration of audio with lecture slide creation was examined in detail. Our system successfully timed the audio recording to the teacher's speech, initiating when the writing commenced and concluding as the board was cleaned. This synchronization was found to be completely aligned and maintained a remarkable tolerance of only 1 second, further demonstrating the system's precision in gathering real-time changes in the classroom.

Comprehensive Evaluation:

During the final step of the evaluation process, we conducted an extensive evaluation of our entire system. The system demonstrated an overall accuracy of 0.89 and an F1 Score of 0.88 in a variety of classroom circumstances as shown in Figure 6. The test results of lecture slide extraction from the classroom board can be further seen in Figure 7. These results confirmed that by precisely recording the instructor's speech and synchronizing it with a generated clear board image, our technology can streamline the creation of lecture materials.

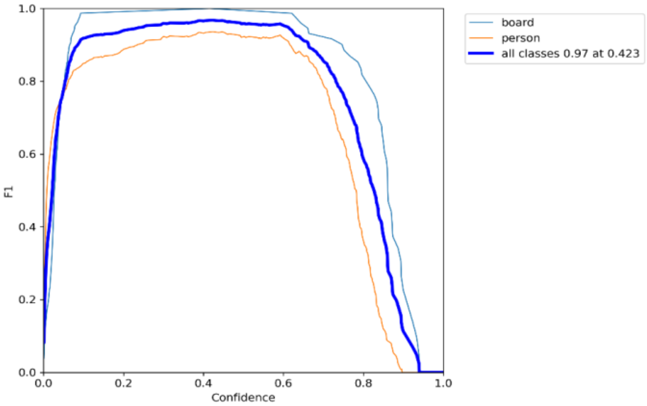

Figure 6. F1 Score for various classes

Discussion:

The integration of YOLO and Scale-Invariant Feature Transform (SIFT) approaches is highlighted as a crucial approach in automating slide extraction from live classroom lectures. By combining these two approaches, our methodology leverages the unique benefits of each strategy to tackle important difficulties faced during the process. YOLO provides a notable benefit with its real-time object-detecting capabilities. This feature quickly identifies important parts in the lecture video stream, such as the board area, teacher, and other relevant things. YOLO's ability to quickly detect objects aids in the initial phase of slide extraction, establishing a solid groundwork for further processing.

Furthermore, using SIFT in this study improved the precision and dependability of slide extraction by implementing a comprehensive feature-based matching technique. SIFT's capability to combine important points from various photos of the same area is extremely valuable for managing differences in lighting, scale, rotation, and viewpoint. This robustness guarantees the accurate retrieval of slide material, even in the presence of distortions or fluctuations in the lecture environment.

Combining YOLO with SIFT allowed for thorough content extraction by utilizing the unique advantages of each method. YOLO is proficient in real-time object identification, while SIFT enhances this ability with precise feature matching and content blending. One of the key strengths of SIFT is its robustness to changes in scale, rotation, and affine transformations. SIFT can accurately identify and match key points even in images with significant variations. This approach is also invariant to changes in illumination and partially invariant to changes in viewpoint, making it a reliable technique for feature extraction. Another strength of SIFT is its ability to extract distinctive and repeatable key points from an image, which can be used for various computer vision tasks such as Object recognition, image retrieval, 3D reconstruction, or image stitching.

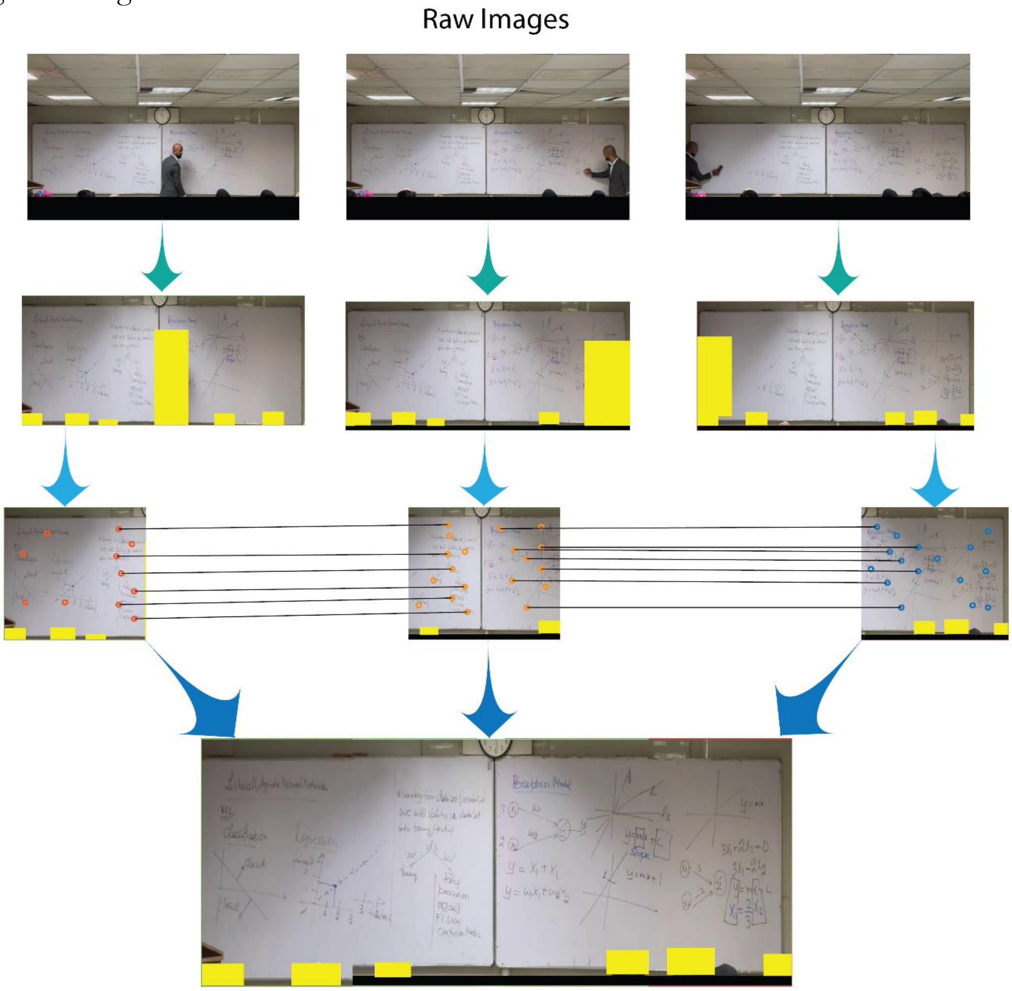

Figure 7. The test results of lecture slide extraction from the classroom board

This synergy leads to more accurate slide extraction, capturing the core of the lecture content with increased precision. The combined technique shows adaptability to various lecture venues, styles, and recording situations. Combining real-time object identification with strong feature-based matching allows efficient slide extraction in various classroom environments. This versatility allows for adjustments in lighting, camera angles, and teacher movements, assuring consistent performance in different situations. The benefits of combining YOLO with SIFT are crucial in improving efficiency and accuracy. The efficient method of extracting slides results in better note-taking decreased manual work in organizing content, and easier access to lecture materials. Consequently, maintaining an enhanced learning experience for pupils is highlighted by the effortless access to thorough and precisely extracted slide material. However, it is important to mention that there might be some overlap in their functionalities of combining both techniques because both YOLO and SIFT are designed to identify objects or features within images. This further leads to redundant processing and additional computational load. Future work could explore techniques to optimize the workflow and minimize redundancy.

Conclusion:

Automated classroom slide extraction using deep learning has considerable potential for improving the productivity and usability of image-based lecture capture systems. By leveraging the complementary strengths of YOLO and SIFT, the proposed method offers enhanced efficiency and accuracy in slide extraction. This results in improved note-taking reduced manual effort in content curation, and enhanced accessibility to lecture materials, ultimately enhancing the overall learning experience for students. Our approach contributes to the broader goal of transforming traditional classroom settings into modern, interactive, and adaptive learning environments. Even though there has been a lot of progress made in this field, there are still a lot of obstacles to overcome in terms of expanding the capabilities of real-time processing, including accuracy and scalability. In the future, research efforts should be directed toward addressing these problems and investigating potential applications of automated slide extraction in educational environments.

Acknowledgment. We acknowledge Dr. Jamil Sawar director of Barani Institute of Technology (BIIT) for support of the dataset and recordings done in this project.

Author's Contribution. All authors contributed equally.

Conflict of interest. There exists no conflict of interest for publishing this manuscript in IJIST.

Project details. This project was conducted as a product for deployment at the Barani Institute of Technology (BIIT) and was completed as a mobile app with a web-based supporting application.

References:

[1] B. S. Prakash, K. V. Sanjeev, R. Prakash, K. Chandrasekaran, M. V. Rathnamma, and V. V. Ramana, "Review of techniques for automatic text summarization," Adv. Intell. Syst. Comput., vol. 1090, pp. 557–565, 2020, doi: 10.1007/978-981-15-1480-7_47/COVER.

[2] M. U. Uçar and E. Özdemir, "Recognizing Students and Detecting Student Engagement with Real-Time Image Processing," Electron. 2022, Vol. 11, Page 1500, vol. 11, no. 9, p. 1500, May 2022, doi: 10.3390/ELECTRONICS11091500.

[3] Z. Chen, M. Liang, Z. Xue, and W. Yu, "STRAN: Student expression recognition based on spatio-temporal residual attention network in classroom teaching videos," Appl. Intell., vol. 53, no. 21, pp. 25310–25329, Nov. 2023, doi: 10.1007/S10489-023-04858-0/METRICS.

[4] G. Lu, "What Factors Influence Students' Seat Preference in the Smart Classroom? Results of Logistic Regression," 2023 11th Int. Conf. Inf. Educ. Technol. ICIET 2023, pp. 426–432, 2023, doi: 10.1109/ICIET56899.2023.10111234.

[5] M. Maijala and M. Mutta, "The Teacher's Role in Robot-assisted Language Learning and its Impact on Classroom Ecology," EuroCALL Rev., vol. 30, no. 2, pp. 6–23, Jan. 2023, doi: 10.4995/EUROCALL.2023.17018.

[6] F. Zulfiqar, R. Raza, M. O. Khan, M. Arif, A. Alvi, and T. Alam, "Augmented Reality and its Applications in Education: A Systematic Survey," IEEE Access, vol. 11, pp. 143250–143271, 2023, doi: 10.1109/ACCESS.2023.3331218.

[7] D. Singh S, A. Gupta, C. V. Jawahar, and M. Tapaswi, "Unsupervised Audio-Visual Lecture Segmentation," Proc. - 2023 IEEE Winter Conf. Appl. Comput. Vision, WACV 2023, pp. 5221–5230, 2023, doi: 10.1109/WACV56688.2023.00520.

[8] S. Kaddoura, S. Vincent, and D. J. Hemanth, "Computational Intelligence and Soft Computing Paradigm for Cheating Detection in Online Examinations," Appl. Comput. Intell. Soft Comput., vol. 2023, 2023, doi: 10.1155/2023/3739975.

[9] E. Dimitriadou and A. Lanitis, "A critical evaluation, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms," Smart Learn. Environ., vol. 10, no. 1, pp. 1–26, Dec. 2023, doi: 10.1186/S40561-023-00231-3/TABLES/7.

[10] C. B. Murthy, M. F. Hashmi, N. D. Bokde, and Z. W. Geem, "Investigations of Object Detection in Images/Videos Using Various Deep Learning Techniques and Embedded Platforms—A Comprehensive Review," Appl. Sci. 2020, Vol. 10, Page 3280, vol. 10, no. 9, p. 3280, May 2020, doi: 10.3390/APP10093280.

[11] J. Prather et al., "The Robots are Here: Navigating the Generative AI Revolution in Computing Education," ITiCSE-WGR 2023 - Proc. 2023 Work. Gr. Reports Innov. Technol. Comput. Sci. Educ., pp. 108–159, Dec. 2023, doi: 10.1145/3623762.3633499.

[12] S. K. Jagatheesaperumal, K. Ahmad, A. Al-Fuqaha, and J. Qadir, "Advancing Education Through Extended Reality and Internet of Everything Enabled Metaverses: Applications, Challenges, and Open Issues," IEEE Trans. Learn. Technol., vol. 17, pp. 1120–1139, 2024, doi: 10.1109/TLT.2024.3358859.

[13] C. Thomas, K. A. V. Puneeth Sarma, S. Swaroop Gajula, and D. B. Jayagopi, "Automatic

prediction of presentation style and student engagement from videos," Comput. Educ. Artif. Intell., vol. 3, p. 100079, Jan. 2022, doi: 10.1016/J.CAEAI.2022.100079.

[14] H. Hassani, M. J. Ershadi, and A. Mohebi, "LVTIA: A new method for keyphrase extraction from scientific video lectures," Inf. Process. Manag., vol. 59, no. 2, p. 102802, Mar. 2022, doi: 10.1016/J.IPM.2021.102802.

[15] F. Kulsoom, S. Narejo, Z. Mehmood, H. N. Chaudhry, A. Butt, and A. K. Bashir, "A review of machine learning-based human activity recognition for diverse applications," Neural Comput. Appl. 2022 3421, vol. 34, no. 21, pp. 18289–18324, Aug. 2022, doi: 10.1007/S00521-022-07665-9.

[16] T. Kalsum et al., "Localization and classification of human facial emotions using local intensity order pattern and shape-based texture features," J. Intell. Fuzzy Syst., vol. 40, no. 5, pp. 9311–9331, Jan. 2021, doi: 10.3233/JIFS-201799.

[17] C. M. Badgujar, A. Poulose, and H. Gan, "Agricultural Object Detection with You Look Only Once (YOLO) Algorithm: A Bibliometric and Systematic Literature Review," Jan. 2024, Accessed: Apr. 14, 2024. [Online]. Available: https://arxiv.org/abs/2401.10379v1

[18] P. Athira, T. P. Mithun Haridas, and M. H. Supriya, "Underwater Object Detection model based on YOLOv3 architecture using Deep Neural Networks," 2021 7th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2021, pp. 40–45, Mar. 2021, doi: 10.1109/ICACCS51430.2021.9441905.

[19] J. Lin, Y. Zhao, S. Wang, and Y. Tang, "YOLO-DA: An Efficient YOLO-Based Detector for Remote Sensing Object Detection," IEEE Geosci. Remote Sens. Lett., vol. 20, 2023, doi: 10.1109/LGRS.2023.3303896.

[20] S. Kim, J. G. Lee, and M. Y. Yi, "Developing information quality assessment framework of presentation slides," https://doi.org/10.1177/0165551516661917, vol. 43, no. 6, pp. 742–768, Sep. 2016, doi: 10.1177/0165551516661917.

[21] I. Benedetto, M. La Quatra, L. Cagliero, L. Canale, and L. Farinetti, "Abstractive video lecture summarization: applications and future prospects," Educ. Inf. Technol., vol. 29, no. 3, pp. 2951–2971, Feb. 2024, doi: 10.1007/S10639-023-11855-W/METRICS.

[22] "Automated analysis and indexing of lecture videos ." Accessed: Apr. 14, 2024. [Online]. Available: https://dr.lib.iastate.edu/server/api/core/bitstreams/d970b9dd-4b3a-4ca8-b75b-645aa693413e/content

[23] T. Seng, "Enriching Existing Educational Video Datasets to Improve Slide Classification and Analysis," MM 2022 - Proc. 30th ACM Int. Conf. Multimed., pp. 6930–6934, Oct. 2022, doi: 10.1145/3503161.3548758.

[24] C. V. Loc, N. T. Nhan, T. X. Viet, T. H. Viet, L. H. Thao, and N. H. Viet, "Content based Lecture Video Retrieval using Textual Queries: to be Smart University," Proc. - Int. Conf. Knowl. Syst. Eng. KSE, vol. 2021-November, 2021, doi: 10.1109/KSE53942.2021.9648820.

[25] B. Teuscher, Z. Xiong, and M. Werner, "An Augmentation Framework for Efficiently Extracting Open Educational Resources from Slideshows," Ninth Int. Conf. High. Educ. Adv., May 2023, Accessed: Apr. 14, 2024. [Online]. Available: http://ocs.editorial.upv.es/index.php/HEAD/HEAd23/paper/view/16169

[26] S. Cumani, P. Laface, and F. Kulsoom, "Speaker recognition by means of acoustic and phonetically informed GMMs," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2015-January, pp. 200–204, 2015, doi: 10.21437/INTERSPEECH.2015-84.